MontCAS, Phase 2
Criterion-Referenced Test

2007
Technical Report

Linda McCulloch, Superintendent
Montana Office of Public Instruction
PO Box 202501
Helena, Montana 59620-2501
www.opi.mt.gov

TABLE OF CONTENTS

SECTION I: ASSESSMENT DEVELOPMENT

CHAPTER 1—BACKGROUND AND OVERVIEW

1.1 PURPOSE OF THIS REPORT

In the spring of 2007, Montana students in grades 3 through 8 and 10 participated in the MontCAS, Phase 2 Criterion Referenced Test (CRT) in reading and mathematics in order to measure their reading and mathematics achievement as articulated by the Montana Content Standards and Grade Level Expectations. This represents the fourth year of the operational CRT program, which was expanded this year to include field tests in science (grades 4, 8 and 10).

The purpose of this report is to describe several technical aspects of the CRT in an effort to contribute to the accumulation of validity evidence to support CRT score interpretations. Because it is the interpretations of test scores that are evaluated for validity, not the test itself, this report presents documentation to substantiate intended interpretations (American Educational Research Association (AERA), American Psychological Association & National Council on Measurement in Education, 1999). Subsequent chapters of this report discuss test development, test alignment, test administration, scoring, equating, item analyses, reliability, scaled scores, performance levels and reporting. Each of these topics contributes important information to the validity of the assessment program. Note however that certain aspects of a comprehensive validity argument are not included in the report that could also be important to consider when drawing conclusions about validity (e.g., additional sources of validity evidence might speak to the extent to which scores from the CRT assessments converge with other measures of the same or similar constructs and diverge from measures of different constructs; consequences that arise from scores at the student, school, district and state levels).

Historically, some parts of technical reports may have been used by educated lay persons, but the intended audience was experts in psychometrics and educational research. This edition of the CRT technical report is an attempt to make the information more accessible to educated lay people, by providing richer descriptions of general categories of information. In making some of the information more accessible, we have purposefully preserved the depth of technical information provided historically. The reader will find that some of the discussion and tables continue to require a working knowledge of measurement concepts such as "reliability" and "validity" and statistical concepts such as "correlation" and "central tendency." To understand fully some of the presented data, the reader will have to possess basic understanding of advanced topics in measurement and statistics.

1.2 Overview of the Assessment System

The CRT was developed in accordance with the following federal laws: Title 1 of the Elementary and Secondary Education Act (ESEA) of 1994, P.L. 103-382 and the No Child Left Behind Act (NCLB) of 2001.

The CRTs are based on, and aligned to, Montana's Content Standards and Grade Level Expectations in Reading and Mathematics. Montana educators worked with OPI and its contractor, Measured Progress, in the development and review (of content and bias) of these tests to assess how well students have learned the Montana content standards for their grade. In addition, an independent alignment study was performed by Northwest Regional Educational Laboratory (NWREL) in fall 2006 prior to test form production for 2007. NWREL's alignment study may be found on OPI's Web site www.opi.mt.gov/assessment.

CRT scores are intended to be useful indicators of the extent to which students have mastered material outlined in the Montana reading and mathematics content standards. For a particular student, his/her CRT score should be used as part of a body of evidence regarding mastery and should not be used in isolation to make high stakes decisions. CRT scores are more reliable indicators of program

success when aggregated to school, system, or state levels, particularly when monitored over the course of several years.

Table 1-1: Timeline of Major Program Milestones

Milestone	Year	Subjects
Montana Content Standards adopted by Montana's Board of Education	1998	Reading and Mathematics
Item development and field test administration of the grades 3 through 8 and 10 CRT Montana-specific items	2003	Reading and Mathematics
First operational administration of the CRT in grades 4, 8 & 10	2004	Reading and Mathematics
Standard Setting for grades 4, 8 and 10	2004	Reading and Mathematics
Second operational administration of the CRT in grades 4, 8 & 10	2005	Reading and Mathematics
Field test administration in grades 3, 5, 6 and 7	2005	Reading and Mathematics
Third operational administration of the CRT in grades 4, 8 & 10; First operational administration of the CRT in grades 3, 5 6 and 7	2006	Reading and Mathematics
Standard Setting for grades 3 through 8 and 10	2006	Reading and Mathematics
Item development and bias review by Montana educators to prepare for science field test in spring 2007	2006	Science
Fourth operational administration of the CRT in grades 4, 8 & 10; Second operational administration of the CRT in grades 3, 5 6 and 7	2007	Reading and Mathematics
Field test administration in grades 4, 8 and 10	2007	Science

1.3 OPTIONS FOR PARTICIPATION

All Montana students enrolled in accredited schools are expected to participate in either the CRT or the CRT Alternate assessment (CRT-ALT). The vast majority of students will participate in the CRT, and most of them will participate under standard administration procedures. However, there is an array of standard accommodations which are available to any student, with or without disabilities, when such accommodations are necessary to allow the student to demonstrate his/her skills and competencies. Standard accommodations are not considered to change the construct being measured and may be provided to students for either the reading or math portions of the assessment, or both, as necessary. Students' tests are scored the same way regardless of whether or not they took the test using standard accommodations.

In addition to standard accommodations, other accommodations for the CRT are available to a student when specified in his/her IEP, 504, or LEP plan. These other accommodations are referred to as non-standard accommodations; because they are considered to alter the construct being measured, they do affect the student's score on the CRT. When a non-standard accommodation is used, the student's score for that content area is reported as the lowest possible (i.e., a scaled score of 200 will fall into the Novice performance level). Non-standard accommodations on the CRT may be provided in reading or math, or both, as dictated by the student's IEP, 504, or LEP plan.

For a very small percentage of students, participation in the statewide assessment program will be achieved by participating in the CRT-ALT. Students with significant cognitive disabilities who are working toward alternate academic achievement standards, as documented in their IEP plans, are eligible to take the CRT-ALT. Technical characteristics of the CRT-ALT program are described in a companion technical report.

1.4 Brief Summary of Technical Evidence in This Report

The *Standards for Educational and Psychological Testing* (AERA et al, 1999) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. These sources include evidence based on the following five general areas: test content, response processes, internal structure, relationship to other variables, and consequences of testing. Although each of these sources may speak to a different *aspect* of validity, they are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

Viewed through this lens provided by the *Standards*, evidence based on test content is extensively described in Chapters 2 through 6. Item alignment with Montana content standards; item bias, sensitivity and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for

participation; and appropriate test administration training are all components of validity evidence based on test content.

The scoring information in Chapter 7 describes the steps taken to train and monitor hand-scorers, as well as quality control procedures related to scanning and machine scoring.

Evidence based on internal structure is presented in detail in the discussions of item analyses and reliability in Chapters 8 and 9. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlations), differential item functioning analyses, standard errors of measurement, dimensionality analyses, and item response theory parameters and procedures, and a variety of reliability coefficients.

Ultimately, the manner in which the test results are reported and used is inextricably related to the concept of validity, and this is addressed in the scale score, equating, and reporting information contained in Chapters 10 and 11, as well as in the test interpretation guide, which is a separate document that is referenced in the discussion of reporting. Each of these chapters speaks to the efforts undertaken to promote accurate and clear information provided to the public regarding test scores.

With this introduction to a conceptual understanding of how the information presented in this report contributes to an overarching validity argument in mind, the reader should be in position to organize the extensive detail contained in the following chapters. The organization of this report is based on the conceptual flow of an assessment cycle. The report begins with the initial test specification and addresses all the intermediate steps that lead to final score reporting.

CHAPTER 2—OVERVIEW OF TEST DESIGN

2.1 CRITERION-REFERENCED TEST (CRT)

Items on the CRT originate from the Measured Progress State Secure Item Pool (MPSSIP) and Montana-augmented item banks (see Chapter 3) and are directly linked to **Montana's Content Standards**. The content standards are the basis for the reporting categories developed for each subject area and are used to help guide the development of test items. No other content or process is subject to statewide assessment. An item may address part, all, or several of the benchmarks within a standard.

2.2 ITEM TYPES

Montana's educators and students were familiar with the item types that were used in the assessment program. The types of items used and the functions of each are described below.

- **Multiple-choice** items were used, in part, to provide breadth of coverage of a content area. Because they require no more than a minute for most students to answer, these items make efficient use of limited testing time and allow coverage of a wide range of knowledge and skills.
- **Short-answer items** were used to assess students' skills and their abilities to work with brief, well-structured problems that had one or a very limited number of solutions (e.g., mathematical computations). Short-answer items require approximately two minutes for most students to answer. The advantage of this type of item is that it requires students to demonstrate knowledge and skills by generating, rather than merely selecting, an answer.
- **Constructed-response items** typically require students to use higher-order thinking skills—evaluation, analysis, summarization, and so on—in constructing a satisfactory response. Constructed-response items should take most students approximately five to ten minutes to

complete. It should be noted that the use of released CRT items to prepare students to answer this kind of item is appropriate and encouraged.

2.3 COMMON-MATRIX DESIGN

The Montana CRTs are structured using both **common** and **field test** items **(**matrix-sampled items.) Common items are those taken by all students at a given grade level. Students' scores are based only on common items. In addition, a larger pool of matrix-sampled items is divided among the sixteen forms of the test at each grade level. Each student takes only one form of the test and so answers a fraction of the matrix-sampled items in the entire pool. The field test items (matrix-sampled items) were transparent to test takers and had a negligible impact on testing time. Because the field test was taken by all students, it provided the sample size needed to produce reliable data (750-1500 students per item) on which to inform item selection for future tests.

The CRT Student reports were delivered to schools on June 29, 2007. All other CRT reporting data were made available to districts and schools online via Measured Progress's secure data management system called iAnalyze. In addition, common items were released on OPI's assessment Web site and on *iAnalyze* (see Chapter 11: "Reporting" and Appendix D: Report Shells.)

CHAPTER 3—TEST DEVELOPMENT PROCESS

3.1 CRITERION-REFERENCED TEST (CRT) ITEM DEVELOPMENT

As previously mentioned, items in the CRT are derived from either the Measured Progress State Secure Item Pool (MPSSIP) or a Montana-augmented item bank. The item development process for both item banks is similar and is discussed in greater detail in this chapter.

3.2 MPSSIP ITEM DEVELOPMENT

The items developed for the Measured Progress State Secure Item Pool (MPSSIP) and forms were consistent with national and Montana Content Standards. Measured Progress curriculum and assessment specialists worked with Montana educators verify the alignment of items to the appropriate Montana Content Standards. As an additional quality control check, Northwest Regional Educational Laboratory (NWREL) performed an independent alignment study to verify item alignment to Montana Content Standards.

The development process that Measured Progress followed combined the expertise of the item development team and a nationwide panel of educators to help ensure that these items met the needs of the core MPSSIP program and the CRT program. All items used in the MPSSIP common portions of the CRT program underwent review by a national panel of content and bias reviewers. This panel included numerous Montana educators. Annual MPSSIP item development is depicted in the following tables:

Table 3-1: Total Number of MPSSIP Items Developed per year Grades 3-8 &10

Grade	Reading	Math
3	160	78
4	160	78
5	160	78
6	160	78
7	160	78
8	160	78
10	160	78

Table 3-2: Annual MPSSIP Reading Item Development Grades 3 - 8 & 10

Passages	Multiple Choice	Constructed Response
2 long literary passages	40	4
2 long informational passages	40	4
4 short literary passages	40	0
4 short informational passages	40	0
12	**160**	**8**

Table 3-3: Annual MPSSIP Math Item Development Grades 3 - 8 & 10

Multiple Choice	Short Answer	Constructed Response
68	4	6

3.3 ITEM DEVELOPMENT PROCESS OVERVIEW

An overview of the test development process for the common and matrix items, including conducting the field tests, follows.

Table 3-4: Development Process Overview

Development Step	Step Details
Select reading passages and conduct external review for bias and sensitivity issues (2006)	Measured Progress Curriculum and Assessment Specialists located potential reading passages. Reading passages were reviewed for bias and sensitivity issues before the development of reading item sets.
Develop items (January through May 2006)	Measured Progress Curriculum and Assessment Specialists developed reading item sets and mathematics items.
National item review for bias and sensitivity issues and content appropriateness (summer 2006)	Panels of national educators reviewed newly-developed reading and mathematics items • to assure items were compliant with the MT bias and sensitivity guidelines and were content appropriate.
Edit items (summer 2006)	All items reviewed by national committee members were edited to assure • clarity and unambiguousness of items • correct grammar, punctuation, usage, and spelling • technical quality with respect to stems, options, and scoring guides.
Item Review and Selection Meeting (summer 2006)	Measured Progress test developers and Montana educators reviewed the results of the Spring 2006 field test and selected common items for the Spring 2007 operational CRT forms.
Montana educators review items for bias and sensitivity issues and content appropriateness (Sept/Oct. 2006)	Panels of Montana educators reviewed reading and mathematics field test items for bias and sensitivity issues and content appropriateness. Montana Educator's editorial comments were incorporated at this time
Field test items (spring 2007)	Embedded matrix (field test) items were administered to a sample of students (minimum of 1,500 students per item/16 forms per grade and content).

3.4 MONTANA-AUGMENTED ITEM DEVELOPMENT

The items developed for the augmented CRT item bank were consistent with Montana's content standards. Using a collaborative model, Measured Progress's development specialists worked with OPI and Montana educators to align the items developed to augment the CRT to appropriate Montana content standards. As an additional quality control check, lead developers in each content area checked for their agreement that each item was appropriately aligned. Where there were any apparent discrepancies, lead Curriculum and Assessment specialists resolved them with OPI personnel.

The development process Measured Progress followed, combining the expertise of the item development team and Montana educators, helped ensure that these items met the needs of the CRT program. The item specifications were built on the Montana content standards, thus assuring complete alignment between the content standards and the augmented portion of the CRT. In addition to internal review, all test materials and items used in the CRT program underwent review by Montana educators and bias review committees prior to print. Table 3-5 depicts the number of items developed and field tested in 2002-2003 to support the program's item bank 2004 through 2007.

Table 3-5: Total Number of Montana-Augmented Items Developed and Field Tested by Grade and Content (all Multiple Choice Items)

Grade	Reading	Math
3	60	60
4	100	100
5	60	60
6	60	60
7	60	60
8	100	100
10	150	150

3.5 MONTANA-AUGMENTED ITEM DEVELOPMENT PROCESS OVERVIEW

The following table presents an overview of the above-described test development process for the Montana-augmented item bank, including conducting the field tests, follows.

Table 3-6: Development Process Overview

Development Step	Step Details
Review by Montana educators of passages for the reading tests (Aug. 2002)	• Measured Progress Curriculum and Assessment reading specialists located potential reading passages. • Montana educators approved the passages in consultation with a Montana Bias Review Committee prior to item writing. • Measured Progress Permissions staff secured permissions to use the passages prior to item writing meetings.
Item drafting/editing meetings (Sept. 2002)	Measured Progress Curriculum and Assessment specialists • provided item development training to Montana participants; • facilitated the development of item ideas by the participants.
Editorial review of items (Oct. 2002)	All items were reviewed by members of Measured Progress's Publications staff to ensure • clarity and unambiguousness of items; • correct grammar, punctuation, usage, and spelling; • technical quality with respect to stems, options, and scoring guides; • compliance with OPI sensitivity standards and style guidelines.
Item review meetings (Nov. 2002)	Curriculum and Assessment Specialists facilitated the review of all items with Montana educators and selected appropriate items for field testing in 2003.
Bias Review Committee meetings (Nov. 2002)	Measured Progress staff facilitated the review of all test items for sensitivity and bias considerations based on OPI guidelines. Members of this committee were selected by OPI. Measured Progress provided OPI with guidelines for committee membership.
Field Test of MT-Augmented Items (April 2003)	Measured Progress provided field test forms which were administered to a sample of students in Montana prior to use of the items in operational assessment to assure quality of items.
Final Item Selection (August 2003)	Measured Progress provided the reports necessary for Montana educators to review the results of field-testing, revise as necessary, and select items for the augmented portion of the assessment.

3.6 INTERNAL ITEM REVIEW

The lead or peer Curriculum and Assessment Specialist within the content specialty reviewed each item for:

- item "integrity", item content and structure, appropriateness to designated content area, item format, clarity, possible ambiguity, keyability, single "keyness", appropriateness and quality of reading selections and graphics, and appropriateness of scoring guide descriptions and distinctions (as correlated to the item and within the guide itself).
- scorability, and evaluated whether the scoring guide adequately addressed performance on the item.
- fundamental issues including the following:
  - What is the item asking?
  - Is the key the only possible key?
  - Is the constructed-response item scorable as written (are the correct words used to elicit the response defined by the guide)?
  - Is the wording of the scoring guide appropriate and parallel to the item wording?

- Is the item complete (i.e., with scoring guide, content codes, key, grade level, and contract identified)?
- Is the item appropriate for the designated grade level?

3.7 External Item and Bias Reviews

All MPSSIP and Montana-augmented items undergo the following external reviews:

- In fall 2006, MPSSIP National Bias and Content Review Committees reviewed common and matrix passages and items used for the 2007 administration during two, two-day meetings, held in Salt Lake City, UT.
- In early December 2006, common item sets were reviewed by Measured Progress content specialists and Montana educators. Feedback from the Montana content and bias reviews were incorporated into the final editing processes.

3.8 Item Editing

Editors reviewed and edited the items to ensure uniform style (based on *The Chicago Report of Style, 15th Edition*) and adherence to sound testing principles. These principles included the stipulation that items

- were correct with regard to grammar, punctuation, usage, and spelling;
- were written in a clear, concise style;
- contained unambiguous explanations for students as to what was required to attain a maximum score;
- were written at a reading level that would allow the student to demonstrate his or her knowledge of the tested subject matter regardless of reading ability;
- exhibited high technical quality regarding psychometric characteristics;
- had appropriate answer options or score-point descriptors; and
- were free of potentially insensitive content.

3.9 OPERATIONAL TEST ASSEMBLY

Test assembly is the sorting and laying out of item sets into test forms. In order to accommodate the embedded field test design, sixteen versions of the test were administered in grade 3 through 8 and 10. Criteria considered during this process included the following:

- **Content coverage/match to test design**. The curriculum specialist completed an initial sorting of items into sets based on a balance of content categories across sessions and forms, as well as a match to the test design (e.g., number of multiple-choice, short-answer, and constructed-response items).
- **Item difficulty and complexity**. Item statistics drawn from the data analysis of previously tested items were used to ensure that there were similar levels of difficulty and complexity across forms.
- **Visual balance**. Item sets were reviewed to ensure that each reflected a similar length and "density" of selected items (e.g., length/complexity of reading selections or number of graphics).
- **Option balance**. Each item set was checked to verify that it contained a roughly equivalent number of key options (As, Bs, Cs, and Ds).
- **Name balance**. Item sets were reviewed to ensure that a diversity of names was used.
- **Bias**. Each item set was reviewed to ensure fairness and balance based on gender, ethnicity, religion, socioeconomic status, and other factors.
- **Page fit**. Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Facing-page issues**. For multiple items associated with a single stimulus (a graphic or a reading selection), consideration was given to whether those items needed to begin on a left- or

right-hand page, as well as to the nature and the amount of material that needed to be placed on facing pages. These considerations served to minimize the amount of page flipping required of the students.

- **Relationships between forms**. Sets of common items were placed identically in each version of the forms. Although matrix-sampled item sets differed from form to form, they took up the same number of pages in each form so that sessions and content areas began on the same page in every form. Therefore, the number of pages needed for the longest form often determined the layout of each form.
- **Visual appeal**. The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of white space, the density of the text, and the number of graphics.

3.10 EDITING DRAFTS OF OPERATIONAL TESTS

Any changes made during the test construction had to be reviewed and approved by the Curriculum and Assessment Specialist. Once a form had been laid out in what was considered its final form, it was reread to identify any final considerations, including the following:

- **Editorial changes**. All text was scrutinized for editorial accuracy, including consistency of instructional language, grammar, spelling, punctuation, and layout. Measured Progress's publishing standards are based on *The Chicago Report of Style, 15th Edition*.
- **Keying items**. Items were reviewed for any information that might "key" or provide information that would help students answer another item. Decisions about moving keying items were based on the severity of the key-in and the placement of the items in relation to each other within the form.
- **Key patterns**. The final sequence of keys was reviewed to ensure that the order appeared random (i.e., no recognizable pattern and no more than three of the same key in a row).

3.11 BRAILLE AND LARGE-PRINT TRANSLATION

Form One for grades 3 through 8, and 10 tests was translated into Braille by National Braille Press, a subcontractor that specializes in test materials for blind and visually impaired students. In addition, *Form One* for each grade was adapted into a large-print version.

CHAPTER 4—DESIGN OF THE READING ASSESSMENT

4.1 READING SPECIFICATIONS

As indicated earlier, the test blueprint/specifications for reading were based on MPSSIP and Montana's reading content standards, which identify five **Montana Content Standards** that apply specifically to reading and reading comprehension. Those content standards follow:

- **Reading Standard 1:** Students construct meaning as they comprehend, interpret, and respond to what they read.
- **Reading Standard 2:** Students apply a range of skills and strategies to read.
- **Reading Standard 3:** Students set goals, monitor, and evaluate their reading progress. (This standard cannot be measured with a traditional paper/pencil test.)
- **Reading Standard 4:** Students select, read, and respond to print and non-print material for a variety of purposes.
- **Reading Standard 5:** Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences.

4.2 ITEM TYPES

The CRT assessments in reading include a mix of multiple-choice and constructed-response items. Constructed-response items required students to write an answer consisting of several phrases or short sentences. Each type of item was worth a specific number of points in the student's total reading score as shown in Table 4-1.

Table 4-1: Item Types

Type of Item	Possible Score Points
Multiple-Choice (MC)	0 or 1
Constructed-Response (CR)	1, 2, 3, or 4

Table 4-2 shows the number of multiple-choice and constructed- response items for grades 3-8 and 10.

Table 4-2:Common Reading Items

				TOTAL	
Grade	**Session 1**	**Session 2**	**Session 3**	**MC**	**CRs**
3 -8	21 MC, 1 CR	10 MC	21 MC, 1 CR	52	2
10	21 MC, 1 CR	15 MC	21 MC, 1 CR	57	2

4.3 DISTRIBUTION OF POINTS ACROSS CONTENT STANDARDS

Table 4-3 shows the distribution of points across content standards.

Table 4.3: Grades 3-8 and Grade 10 Reading Specifications/Blueprint

Number of Points on the Common (Scored) Test:	**Grades 3-8 : 52 MC items + 2 CR items = 60 points** **Grade 10: 57 MC items + 2 CR items = 65 points**						
Percent Point distribution by content standard*							
Montana Content Standards	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**	**Grade 10**
Standard 1	34%	34%	34%	34%	34%	34%	25%
Standard 2	30%	30%	30%	30%	30%	30%	32%
Standard 3							
Standard 4	18%	18%	18%	18%	18%	18%	22%
Standard 5	18%	18%	18%	18%	18%	18%	22%
*Because percents are rounded to the nearest whole number, not all sums add to 100%. **Note:** Standard 3 cannot be measured with a traditional paper/pencil test.							
Target point distribution by content standard (Acceptable Range)							
Montana Content Standards	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**	**Grade 10**
Standard 1	20 (18-22)	20 (18-22)	20 (18-22)	20 (18-22)	20 (18-22)	20 (18-22)	16 (14-18)
Standard 2	18 (16-20)	18 (16-20)	18 (16-20)	18 (16-20)	18 (16-20)	18 (16-20)	20 (18-22)
Standard 3							
Standard 4	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	14 (12-16)
Standard 5	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	11 (9-13)	14 (12-16)
Four-point items: Each test contains two 4-point constructed-response items. In any given year, the two items will measure two different standards. From year to year, those standards may change.							
One-point items: The number of one-point items per content standard will vary from year to year depending on which two standards are measured by the four-point items. (The number of total points per standard falls within the acceptable range from year to year.)							

4.4 PASSAGE TYPES

Passages included both long and short texts selected from reading sources that students at each grade level would be likely to encounter in their classroom or in their independent reading. No passages were written specifically for the assessment, but instead were collected from published works. Each passage is classified as one of three types described below.

- **Literary passages** are represented by a variety of genres—modern narratives; diary entries; drama; poetry; biographies; essays; excerpts from novels; short stories; and traditional narratives, such as fables, myths, and folktales.
- **Content passages** are primarily informational and often deal with the areas of science and social studies. They are drawn from such sources as newspapers, magazines, and books.
- **Practical passages** are functional materials that instruct or advise the reader—for example, directions, reference tools, or reports.

The main difference in the passages used for grades 3 – 8, and 10 was their degree of difficulty. All passages were selected to be appropriate for the intended audience; however, the ideas expressed became increasingly more complex from grades 3 through grade 10.

The items related to these passages required students to demonstrate their skills in both literal comprehension, where the answer is stated explicitly in the text, and inferential comprehension, where the answer is implied by the text and/or the text must be connected to relevant prior knowledge to determine an answer. In addition, some items focused on the reading skills reflected in content standards. Items of this type required students to use reading skills and strategies to answer items—for example, how to identify the author's principal purpose, such as to persuade, entertain, or inform—and to demonstrate their understanding of how words and images communicate to readers. Tables 4-4 & 4-5 depict passage distribution and length for Grades 3-8 and Grade 10.

Table 4-4 Passage Distribution Grades 3-8*

Reading Passage Distribution			
Literary	Stories, poetry, and other forms of literature	50 %	30 points
Informational	Content and practical passages	50 %	30 points
		100 %	60 points
Reading Passage Length			
Long	One literary or one informational per session	50 %	30 points
Short	At least one literary and informational per session	50 %	30 points
		100 %	60 points

Table 4-5 Passage Distribution Grade 10*

Reading Passage Distribution			
Literary	Stories, poetry, and other forms of literature	50 %	33 points
Informational	Content and practical passages	50 %	32 points
		100 %	65 points
Reading Passage Length			
Long	One literary or one informational per session	50 %	33 points
Short	At least one literary and informational per session	50 %	32 points
		100 %	65 points

* an example of a generic scoring rubric can be found in table 7-10

While every attempt is made to adhere to recommended grade-level word counts for long and short passages, the final decision in the passage selection process is based on extensive reviews by content experts and bias panels, as well as a careful analysis of the sophistication of language, complexity of concepts, and readability of each passage. Table 4-6 shows the approximate length of the passages selected for the CRT.

Table 4-6 Approximate Length of Passages

Grade Level	Long Passage (number of words)*	Short Passage (maximum word length)*
Grade 3	350-800	350
Grade 4	400-850	400
Grade 5	450-850	450
Grade 6	450-900	450
Grade 7	450-950	450
Grade 8	500-1,000	500
Grade 10	550-1,200	550

Table 4-7: Grades 3-8 Reading Test Design With Field Test Items

Passages	Number of Items
Session 1: Common	
Short passage A	5 MC
Short passage B	5 MC
Long passage A	11 MC, 1 CR
Session 1 Total	***21 MC, 1 CR***
Session 2: Common Augmented & Embedded Matrix (field test) Items	
Augmented Passages	10 MC (common)
Embedded Short Passage	5 MC (field test items)
Embedded Long Passage	7 MC, 1 CR (field test items)
Session 2 Total	***22 MC, 1 CR***
Session 3: Common	
Short passage C	5 MC
Short passage D	5 MC
Long passage B	11 MC, 1 CR
Session 3 Total	***21 MC, 1 CR***
Common (Scored) Total	***52 MC, 2 CR***
Test Total	**64 MC, 3 CR**

Table 4-8: Grade 10 Reading Test Design With Field Test Items

Passages	Number of Items
Session 1: Common	
Short passage A	5 MC
Short passage B	5 MC
Long passage A	11 MC, 1 CR
Session 1 Total	***21 MC, 1 CR***
Session 2: Common Augmented & Embedded Matrix (field test) Items	
Augmented Passages	15 MC (common)
Embedded Short Passage	5 MC (field test items)
Embedded Long Passage	7 MC, 1 CR (field test items)
Session 2 Total	***27 MC, 1 CR***
Session 3: Common	
Short passage C	5 MC
Short passage D	5 MC
Long passage B	11 MC, 1 CR
Session 3 Total	***21 MC, 1 CR***
Common (Scored) Total	***57 MC, 2 CR***
Test Total	**69 MC, 3 CR**

CHAPTER 5—DESIGN OF THE MATHEMATICS ASSESSMENT

5.1 MATHEMATICS SPECIFICATIONS

Mathematics specifications/blueprint is based on Montana's Mathematics Content Standards, which identifies seven standards:

- **Mathematics Standard 1:** Problem Solving
- **Mathematics Standard 2:** Numbers and Operations
- **Mathematics Standard 3:** Algebra
- **Mathematics Standard 4:** Geometry
- **Mathematics Standard 5:** Measurement
- **Mathematics Standard 6:** Data Analysis, Statistics, and Probability
- **Mathematics Standard 7:** Patterns, Relations, and Functions

Table 5-1: Mathematics Specifications/Blueprint

Test Design:	45 multiple-choice items 3 1-point short-answer items 2 4-point constructed-response items Total points: 56

Percent Point distribution by content strand*

MPSSIP Standards	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Number and Operations	32%	32%	32%	32%	30%	20%	20%
Algebra	20%	20%	20%	20%	20%	29%	27%
Geometry	16%	16%	16%	16%	16%	18%	23%
Measurement	13%	13%	13%	13%	14%	14%	11%
Data Analysis/Probability	20%	20%	20%	20%	20%	20%	20%

*Because percents are rounded to the nearest whole number, not all sums add to 100%.

Note: Geometry and Measurement comprise a single reporting category.

Point distribution by content strand

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Number and Operations	18	18	18	18	17	11	11
Algebra	11	11	11	11	11	16	15
Geometry	9	9	9	9	9	10	13
Measurement	7	7	7	7	8	8	6
Data Analysis/Probability	11	11	11	11	11	11	11

Four-point items: Each test contains two 4-point constructed-response items. In any given year, the two items will measure two different strands. From year to year, those strands may change.

One-point items: There are two types of one-point items: multiple-choice and short answer items. Each test contains 45 multiple-choice items and three short-answer items. The number of one-point items per strand will vary from year to year depending on which two strands are measured by the four-point items. (The number of total points per strand is kept constant from year to year.)

Number of 1-point items per content strand

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Number and Operations	14 or 18	14 or 18	14 or 18	14 or 18	13 or 17	7 or 11	7 or 11
Algebra	7 or 11	7 or 11	7 or 11	7 or 11	7 or 11	12 or 16	11 or 15
Geometry	5 or 9	5 or 9	5 or 9	5 or 9	5 or 9	6 or 10	9 or 13
Measurement	3 or 7	3 or 7	3 or 7	3 or 7	4 or 8	4 or 8	2 or 6
Data Analysis/Probability	7 or 11	7 or 11	7 or 11	7 or 11	7 or 11	7 or 11	7 or 11

Distribution of One-Point Items Within Strand by Standard

The distribution of one-point items within a strand is partially dependent on the specific items selected for a given test. However, a minimal number of one-point items per standard have been established. Those numbers are shown in the table below.

Minimum Number of 1-Point Items Per Strand

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
Number and Operations							
Total Number of points	***18***	***18***	***18***	***18***	***17***	***11***	***11***
Number concepts	4	3	2	3	3	2	2
Meanings of operations	1	1	1	1	1	1	1
Computation/estimation	4	5	6	5	4	2	2
Floating points	*5 or 9*	*5 or 9*	*5 or 9*	*5 or 9*	*5 or 9*	*2 or 6*	*2 or 6*
Algebra							

Total Number of points	***11***	***11***	***11***	***11***	***11***	***16***	***15***
Patterns	3	2	2	1	1	1	1
Algebraic symbols	1	1	1	2	2	4	4
Mathematical models	1	1	1	1	1	1	1
Change	1	1	1	1	1	1	1
Floating points	*1 or 5*	*2 or 6*	*2 or 6*	*2 or 6*	*2 or 6*	*5 or 9*	*4 or 8*
Geometry							
Total Number of points	***9***	***9***	***9***	***9***	***9***	***10***	***13***
Properties of 2-and 3-d shapes	2	2	2	2	2	2	3
Coordinate Geometry	1	1	1	1	1	1	1
Transformations/symmetry	1	1	1	1	1	1	1
Visualization/spatial reasoning	1	1	1	1	1	1	1
Floating points	*0 or 4*	*0 or 4*	*0 or 4*	*0 or 4*	*0 or 4*	*1 or 5*	*3 or 7*
Measurement							
Total Number of points	***7***	***7***	***7***	***7***	***8***	***8***	***6***
Concepts of measurement	1	1	1	1	1	1	1
Techniques, tools, formulas	1	1	1	1	1	1	1
Floating points	*1 or 5*	*1 or 5*	*1 or 5*	*1 or 5*	*2 or 6*	*2 or 6*	*0 or 4*
Data Analysis/Probability							
Total Number of points	**11**	**11**	**11**	**11**	**11**	**11**	**11**
Collect/organize/display data	2	2	2	1	1	1	1
Statistical methods	1	1	1	1	1	1	1
Inferences/predictions	1	1	1	1	1	1	1
Probability	1	1	1	1	1	1	1
Floating points	*2 or 6*	*2 or 6*	*2 or 6*	*3 or 7*	*3 or 7*	*3 or 7*	*3 or 7*

5.2 CONTENT SPECIFICATIONS

For students to function effectively as mathematical problem solvers, they must be taught how to apply and communicate basic concepts and procedures, as well as how to do the procedures themselves.

- **Content items** measure what students have been taught. Included in these are the basic concepts and procedural skills from all the content standards. For example, in the numbers and number sense standard and the computation standard, conceptual and procedural knowledge includes understanding of place value in the number system; the computational algorithms as applied to whole numbers, fractions, and decimals; and the concepts of ratio, proportion, and percent. In the data analysis and statistics standard, conceptual and procedural knowledge includes the ability to read charts and graphs as well as to understand concepts of averages

(means, medians, and modes) and the methods for computing them. Contextual settings used in items measuring this category were very simple and were directly related to those used in the teaching of the concepts and the procedures.

- **Application items** measure what the students can do with the content they have learned. Included are items requiring students to combine the basic concepts and procedures to solve real-life and mathematical problems, to evaluate their own ideas and the ideas of others using mathematical reasoning, and to communicate their ideas using the wealth of symbolic, pictorial, graphic, and verbal representations available in mathematics.

It is important to understand that application items also measure mastery of the basic concepts and procedures. For example, in mathematics, items were either short-answer or constructed-response items (see "Item Types" in the table below), which were worth up to four score points. In most cases, portions of these items required the student to perform some problem solving, reasoning, and/or communicating. At the same time, however, the items required the students to demonstrate their understanding of mathematics content. If a student did not show mastery of all aspects of a constructed-response item, or if he/she made careless errors, the student did not earn the highest score for that item. Thus, it can be said that **all** mathematics items in the CRT measured content; some items went beyond that realm (short-answer and constructed-response), however, and were classified as application.

Table 5-2: Distribution of Mathematics Process Categories

Grade	3	4	5	6	7	8	HS
Basic Concepts/ Procedures	65%	65%	60%	60%	55%	55%	55%
Problem Solving/ Reasoning	35%	35%	40%	40%	45%	45%	45%

5.3 ITEM TYPES

The CRT mathematics assessment included multiple-choice, short-answer, and constructed-response items. Short-answer items required students to perform a computation or solve a simple problem. Constructed-response items were more complex, requiring 8-10 minutes of response time. Each type of item was worth a specific number of points in the student's total mathematics score, as shown below.

Table 5-3: Item Types

Type of Item	Possible Score Points*
Multiple-Choice	0 or 1
Short-Answer	0 or 1
Constructed-Response	0, 1, 2, 3, or 4

* an example of a generic scoring rubric can be found in table 7-10

5.4 TEST DESIGN

Table 5-4 summarizes the number and types of items that were used in the CRT mathematics assessment for 2007, and shows the construction of the common portions of the assessment.

Table 5-4: Common Mathematics Items

					TOTAL	
Grade	**Session 1 Cal**	**Session 2A Cal**	**Session 2B No Cal**	**Session 3 No Cal**	**MC**	**SA & CRs**
3	24 MC, 1 CR	5 MC	5 MC	21 MC, 3 SA, 1 CR	55	3 SA, 2 CRs
4	24 MC, 1 CR	5 MC	5 MC	21 MC, 3 SA, 1 CR	55	3 SA, 2 CRs
5	24 MC, 1 CR	5 MC	5 MC	21 MC, 3 SA, 1 CR	55	3 SA, 2 CRs
6	24 MC, 1 CR	5 MC	5 MC	21 MC, 3 SA, 1 CR	55	3 SA, 2 CRs
7	24 MC, 1 CR	5 MC	5 MC	21 MC, 3 SA, 1 CR	55	3 SA, 2 CRs
8	24 MC, 1 CR	5 MC	5 MC	21 MC, 3 SA, 1 CR	55	3 SA, 2 CRs
10	24 MC, 1 CR	8 MC	7 MC	21 MC, 3 SA, 1 CR	60	3 SA, 2 CRs

Cal = calculator use allowed **No Cal** = no calculator use allowed **MC** = multiple-choice items **SA** = short-answer items **CR** = constructed-response items

5.5 THE USE OF CALCULATORS IN THE CRT

The Montana educators who helped develop the CRT acknowledged the importance of mastering arithmetic algorithms. At the same time, they understood that the use of calculators is a necessary and important skill in society today. Calculators can save time and prevent error in the

measurement of some higher-order thinking skills and allow students to do more sophisticated and intricate problems. For these reasons, calculators were permitted on some parts of the CRT mathematics assessment and prohibited on others. (Students were allowed to use any calculator with which they were familiar.)

SECTION II: TEST ADMINISTRATION

CHAPTER 6—TEST ADMINISTRATION

6.1 RESPONSIBILITY FOR ADMINISTRATION

As indicated in the *Test Coordinator's Manual*, principals and/or their designated School Test Coordinators were responsible for the proper administration of the CRT. This report was used to ensure the uniformity of administration procedures from school to school.

6.2 PROCEDURES

School Test Coordinators were instructed to read the *Test Coordinator's Manual* prior to testing, and to be familiar with the instructions given in the *Test Administrator's Manual*. The *Test Coordinator's Manual* provided each school with checklists to help prepare for testing. The checklists outlined tasks to be performed before, during, and after test administration. Along with providing these checklists, the *Test Coordinator's Manual* outlined the nature of the testing material being sent to each school, how to inventory the material, how to track it during administration, and how to return the material once testing was complete. It also contained information about including or excluding students. The *Test Administrator's Manual* included checklists for the administrators to prepare themselves, their classrooms, and their students for the administration of the test. The *Test Administrator's Manual* contained sections that detailed the procedure to be followed for each test session, and it contained instructions on preparing the material prior to giving it to the School Test Coordinator for its return to Measured Progress.

6.3 TEST ADMINISTRATOR TRAINING

In addition to distributing the *2007 Test Coordinator's Manuals* and *Test Administrator's Manuals*, OPI and Measured Progress produced and distributed two audio PowerPoint presentations, "Spring 2007: CRT and CRT-ALT Overview and Update of System and School Test Coordinators"

and "CRT-ALT Test Administrator Training CD" to each system and school test coordinator. Training materials and the audio PowerPoint presentations were also posted on OPI's Web site. System and school test coordinators were not required to travel long distances to attend pre-administration workshops and they could share the training CD with other educators within their buildings.

6.4 PARTICIPATION REQUIREMENTS

All students were expected to participate in the CRT; however, the scores of students in the following categories were excluded from the calculation of averages:

- Foreign exchange students
- Students not enrolled in an accredited Montana school (for example: home-schooled student)
- Students enrolled in a private accredited school
- Students enrolled in a private non-accredited school
- Students enrolled in a private non-accredited Title 1 school
- Students enrolled part-time (less than 180 hours) taking a mathematics or reading course
- First year in US LEP students **were required** to participate in the math assessment only.
- Student took the CRT using a "non-standard" accommodation.

A summary of this information is shown in the table below which was published in the *Test Administrator's Manual* and *Test Coordinator's Manual*.

Table 6-1: Summary of Eligibility for Exclusion from the CRT

Excluded from averages	MUST Participate	MAY Participate
Foreign exchange students	Yes	
Students not enrolled in an accredited Montana school		Yes
Students enrolled in a private accredited school	Yes	
Students enrolled in a private non-accredited school		Yes
Students enrolled in a private non-accredited Title I school		Yes
Students enrolled part-time (less than 180 hrs.) taking a mathematics or reading course		Yes
Reading: first year in US LEP students		Yes
Mathematics: First year in US LEP students	Yes	

Information about the exclusion was coded in by staff after testing was completed in the Student Response Booklet, if applicable. The *Test Coordinator's Manual* and *Test Administrator's Manual* provided detailed instructions for coding exclusions and accommodations. In addition, testing exclusions were discussed thoroughly in the pre-administration training audio CD. Please refer to Appendix E: Reporting Decision Rules.

6.5 TEST SCHEDULING

The CRTs were given during the spring: **reading** and **mathematics** were administered to grades 3 through 8 and 10 during the four-week period, March 6–29, 2007. Schools were able to schedule testing sessions at any time during this period, provided they followed the sequence in the scheduling guidelines detailed in *Test Administrator's Manual.* Schools were asked to schedule makeup testing of students who were absent from initial test sessions during this testing window.

The CRT is an un-timed assessment; however, guidelines or ranges were provided in the *2007 Test Coordinator's Manual* and *2007 Test Administrator's Manual* based on estimates of the time it would take an average student to respond to each type of item that made up the test:

- multiple-choice items – 1 minute per item
- short-answer items – 2 minutes per item
- constructed-response items – 10 minutes per item

While the guidelines for scheduling were based on the assumption that most students would complete the test within the time estimated, each test administrator was asked to allow additional time for students who needed it (see Tables 6-2 through 6-5). If additional classroom space was not available for students who required additional time to complete the tests, schools were encouraged to consider using another space, such as the guidance office, for this purpose. If additional areas were not available, it was recommended that each classroom being used for test administration be scheduled for the maximum amount of time.

Table 6-2: Grades 3 through 8 Recommended Testing Schedule for Reading

DAY 1 Reading	Test Activity	Time Range (in minutes)
	General Instructions	5-10
Session 1	Reading Session 1	45-55
DAY 2		
Session 2	Reading Session 2	45-55
	Break	
Session 3	Reading Session 3	45–55

Table 6-3: Grades 3 through 8 Recommended Testing Schedule for Mathematics

DAY 3 Mathematics	Calculators ARE allowed	Time Range (in minutes)
Session 1	Mathematics Session 1	45-55
	Break	
Session 2A	Mathematics Session 2A	20-30
DAY 4 Mathematics	**Calculators are NOT allowed**	
Session 2B	Mathematics Session 2B	20-30
	Break	
Session 3	Mathematics Session 2B	45-55

TABLE 6-4: GRADE 10 Recommended TESTING SCHEDULE FOR READING

DAY 1 Reading	Test Activity	Time Range (in minutes)
	General Instructions	10–20
Session 1	Reading Session 1	50–60
DAY 2 Reading		
Session 2	Reading Session 2	50–60
	Break	
Session 3	Reading Session 3	50–60

Table 6-5: Grade 10 Recommended Testing Schedule for Mathematics

DAY 1 Mathematics	Calculators ARE allowed	Time Range (in minutes)
Session 1	Mathematics Session 1	50–60
	Break	
Session 2A	Mathematics Session 2A	20–30
DAY 2 Mathematics	**Calculators are NOT allowed**	
Session 2B	Mathematics Session 2B	20–30
	Break	
Session 3	Mathematics Session 3	50–60

6.6 HELP DESK

To address testing concerns, Measured Progress established a help desk dedicated to the State of Montana. Help desk support is an essential element to the successful administration of large-scale assessments. It provides a centralized location where individuals in the field can call a toll-free number to request assistance, report problems they are experiencing, or ask specific questions.

The Measured Progress help desk provided support during all phases of the testing window. It was staffed at varying levels based on need and volume and was available from 8:00 A.M. to 4:00 P.M. MST during the testing window. At a minimum, the help desk consisted of a product support specialist who was responsible for receiving, responding to, and tracking calls and e-mails, and routing issues to the appropriate person(s) for resolution. In addition, communications requiring a higher level of program support were routed to the program manager and/or program assistant

When possible, all calls and e-mails received during business hours were responded to immediately with resolution or updated within hours of receipt.

SECTION III: DEVELOPMENT AND REPORTING OF SCORES

CHAPTER 7—SCORING

Scoring of multiple-choice, short-answer, and constructed-response items is the most important process of any large-scale assessment. The following paragraphs define the scoring processes used for Montana's Criterion-Referenced Test (CRT) program.

7.1 SCANNING

Months prior to test administration and subsequent scanning activities, the scanning department met with the program management team to determine decision rules and required scanning and imaging specifications. The information gathered at these meetings was then used to develop a customized scanning program for Montana.

For the Montcas CRT program Measured Progress used the NCS 5000i scanners, which offer rapid, highly accurate scanning and imaging technology. The 5000i scanners feature numerous real-time quality control checks, such as duplex read, a transport printer that prints a unique identifying number on each sheet of each booklet, and on-line editing capability,

At the conclusion of testing, Montana schools shipped all test materials back to Measured Progress. To expedite the scanning and scoring process, used student response booklets were express-shipped separately from other test materials. Once the 77,459 used student response booklets were logged in, identified with appropriate scannable, preprinted school information sheets, examined for extraneous materials, and batched, they were moved into the scanning area.

The first step in that conversion was the removal of the booklet bindings so that the individual pages could pass through the scanners one at a time. Once cut, the sheets were put back in their proper boxes and placed in storage until needed for the scanning/imaging process.

Customized scanning programs for all scannables were prepared to selectively read the student response booklets and to format the scanned information electronically according to predetermined requirements. Any information (including multiple-choice response data) that had been designated time-critical or process-critical was handled first.

All student response documents and other scannable information necessary to produce the required reports were captured and converted into an electronic format, including all student identification and demographics, and digital image clips of short-answer and constructed-response student responses. The digital image clip information allowed Measured Progress to replicate student responses on the readers' monitors just as they had appeared on the originals. From that point on, the entire process—data processing, scoring, benchmarking data analysis, and reporting—was accomplished without further reference to the originals.

7.2 Scanning Quality Control

The scanners are equipped with many built-in safeguards that prevent data errors. The scanning hardware is continually monitored for conditions that cause the machine to shut down if standards are not met. It will display an error message and prevent further scanning until the condition is corrected. The areas monitored include document page and integrity checks, user-designed on-line edits, and many internal checks of electronic functions.

In an effort to protect data integrity Measured Progress operators perform a diagnostic routine before every scanning shift begins. In the rare event that the routine detects a photocell that appears to be out of range, that machine is re-calibrated and tested again. If the read is still not up to standard, field service engineer is called in for assistance.

As a final safeguard, spot checks of scanned files, bubble by bubble and image by image, were routinely made throughout scanning runs. The result of these precautions, from the original layout of the scanning form to the daily vigilance of the operators, was a scan error rate well below 1 per 1000.

7.3 Electronic Data Files

Once the scanning process was completed, the booklets themselves were put into storage (where they stayed for at least 180 days beyond the close of the fiscal year). When it had been determined that the files were complete and accurate, those files were duplicated electronically and made available for many other processing options. Completed files were loaded onto the local area network (LAN) for transfer to Measured Progress's proprietary *iScore* system for scoring. Those files were then used to identify (and print out) papers to be used in the benchmarking processes, and the data made transferable via the Internet, CD-ROM, or optical disk.

Table 7-1: Number of Responses Scanned and Scored

Grade/Content	Number of Responses Scanned and Scored
3 Math	83,922
4 Math	82,885
5 Math	84,090
6 Math	85,662
7 Math	90,250
8 Math	91,621
10 Math	93,058
3 Reading	31,382
4 Reading	31,322
5 Reading	31,792
6 Reading	32,442
7 Reading	33,727
8 Reading	34,655
10 Reading	35,210

7.4 Items Scored by Readers

Test and answer materials were handled as little as possible to minimize the possibility of loss, mishandling, or breach of security. Once scanned, either by optical mark reader or the *iScore* system, papers were stored securely in areas with limited personnel access.

As explained in the following sections on scoring, the *iScore* system itself ensures the security of responses and test items: all scoring is "blind"; that is, no student names are associated with viewed responses or raw scores and all scoring personnel are subject to the same nondisclosure requirements and supervision as regular Measured Progress staff.

7.5 iSCORE

All of Measured Progress's scoring facilities use the iScore process. iScore is Measured Progress's Web-based proprietary software used to score short-answer and constructed response items. Images of student responses are transferred electronically via a secure Web site to a scorer's computer screen at any one of Measured Progress's scoring facilities. For Montana's CRT program, scoring took place in Dover, New Hampshire, Albany, New York, Denver, Colorado and Louisville, KY.

After the 2007 test material had been loaded into the LAN, *iScore* sent electronically scanned images of student work to individual readers at computer terminals, who evaluated each response and recorded each student's score via keypad or mouse entry. When the reader had finished with one response, the next response appeared immediately on the computer screen. In that way, the system guaranteed complete anonymity of individual students and ensured the randomization of responses during scoring.

Although *iScore* is based on conventional scoring techniques, it also offers numerous benefits:

- real-time information on scorer reliability, read-behinds, and overall process monitoring;
- early access to subsets of data for tasks such as standard setting;
- reduced material handling, which saves time and labor and enhances the security of materials; and
- immediate access to samples of student responses and scores for reporting and analysis through electronic media.

Scoring operations, directed by the manager of scoring services, were carried out by a highly qualified staff. The staff included

- chief readers, who oversaw all training and scoring within particular subject areas;
- quality assurance coordinators (QACs), who led benchmarking and training activities and monitored scoring consistency and rates;
- verifiers, who performed read-behinds of readers and assisted at scoring tables as necessary; and
- readers, who performed the bulk of the scoring.

Table 7-2 summarizes the qualifications of the 2007 CRT quality assurance coordinators and readers.

Table 7-2: Educational Credentials

Montana Reader Education Credentials

Description	Albany, NY	Denver, CO	Dover, NH	Louisville, KY	Total	Pct
Less then 48 college credits	0	0	0	0	0	0.00%
48+ college credits	7	0	0	3	10	3.44%
Associate's degree	6	0	1	6	13	4.47%
Bachelor's degree	52	10	10	108	180	61.86%
Master's degree	27	1	5	38	71	24.40%
Doctorate	6	1	0	10	17	5.84%
Total	**98**	**12**	**16**	**165**	**291**	

Montana QAC Education Credentials

Description	Albany, NY	Denver, CO	Dover, NH	Louisville, KY	Total	Pct
Less then 48 college credits	0	0	0	0	0	0.00%
48+ college credits	1	0	0	0	1	1.72%
Associate's degree	0	0	0	1	1	1.72%
Bachelor's degree	8	3	4	19	34	58.62%
Master's degree	4	2	3	11	20	34.48%
Doctorate	0	0	0	2	2	3.45%
Total	**13**	**5**	**7**	**33**	**58**	

7.6 Preliminary Activities

The preliminary activities for scoring included participating in the planning and design of documents to be used for scoring, reviewing items and score guides for benchmarking and training and the creation of benchmarking packets, and selecting scoring staff and training them for scoring.

7.7 Planning and Designing Documents

At the request of the project manager, scoring personnel advised project management and OPI staff on the program design in order to support an efficient and effective scoring process. Scoring staff also contributed to the design of

- response documents and the image-capture process to yield acceptable image clips (also defining file format and layout); and
- scoring benchmarks composed of the guide, subject background information, and anchor papers.

7.8 Benchmarking

Before the scheduled start of scoring activities, scoring center staff and Montana educators reviewed test items and scoring guides for benchmarking. At that point, chief readers and selected QACs prepared scorer training materials.

Scoring staff from Measured Progress (including test developers) and Montana educators selected one or two anchor examples for each item score point. An additional six to ten responses per item were chosen as part of the training pack. The anchor pack consisted of midrange exemplars, while the training pack exemplars illustrated the range within each score point. The chief readers, who worked closely with QACs for each content area, facilitated the selection of response exemplars.

7.9 Selecting and Training Scoring Staff

Quality Assurance Coordinators (QACs) and Verifiers

Because the read-behinds performed by the QACs and verifiers moderated the scoring process and thus maintained the integrity of the scores, individuals chosen to fill those positions were selected for their accuracy.

In addition, QACs, who train readers to score each item in their content areas, were selected for their ability to instruct and for their level of expertise in their content areas. For this reason, QACs typically are retired teachers who have demonstrated a high level of expertise in their respective disciplines. The ratio of QACs and verifiers to readers was approximately 1:11.

Training Quality Assurance Coordinators and Verifiers

To ensure that all QACs provided consistent training and feedback, the chief readers spent two days training and qualifying the QACs, and the QACs reviewed all items with the verifiers before scoring. In addition, QACs rotated among tables, supervising readers and reading behind verifiers, who in turn read behind a different table of readers each day.

Selecting Readers

Applicants were required to demonstrate their ability by participating in a preliminary scoring evaluation. The *iScore* system enables Measured Progress to efficiently measure a prospective reader's ability to score student responses accurately. After participating in a training session, applicants were required to achieve at least 80% exact scoring agreement for a qualifying pack consisting of 20 responses to a predetermined item in their content area. Those 20 responses were randomly selected from a bank of approximately 150, all of which had been selected by QACs and approved by the chief readers and developers. Table 7-3 depicts the accuracy and qualification percentages of the readers.

Table 7-3: Montana Scoring Accuracy and Qualification Statistics 2007

Content	Grade	Item	Average % Exact Agreement for Embedded CR sets	Average % Exact Agreement for Double Blind Scoring	Number of Readers taking Qualification Sets	Number Successfully Qualifying	Percent Successfully Qualifying
Math	3	25	93.0	94.8	11	10	90.9
		65	71.2	98.3	NA	NA	NA
		66	NA	97.6	NA	NA	NA
		67	NA	97.1	NA	NA	NA
		68	90.0	79.1	12	9	75.0
Math	4	25	94.7	94.8	20	14	70.0
		65	NA	98.3	NA	NA	NA
		66	NA	97.6	NA	NA	NA
		67	NA	97.1	NA	NA	NA
		68	92.1	79.1	20	17	85.0
Math	5	25	86.4	88.4	21	19	90.5
		65	NA	88.7	NA	NA	NA
		66	NA	96.9	NA	NA	NA
		67	NA	97.7	NA	NA	NA
		68	89.5	89.9	21	14	66.7
Math	6	25	78.0	81.1	25	17	68.0
		65	NA	89.8	NA	NA	NA
		66	NA	94.2	NA	NA	NA
		67	NA	96.9	NA	NA	NA
		68	74.3	81.4	21	11	52.4
Math	7	25	91.1	89.4	15	14	93.3
		65	NA	96.8	NA	NA	NA
		66	NA	96.8	NA	NA	NA
		67	NA	94.0	NA	NA	NA
		68	85.9	92.3	16	15	93.8
Math	8	25	94.2	72.0	14	13	92.9
		65	NA	94.4	NA	NA	NA
		66	NA	93.3	NA	NA	NA
		67	NA	97.8	NA	NA	NA
		68	86.5	94.3	14	13	92.9
Math	10	25	87.0	92.2	15	12	75.0
		70	NA	96.3	NA	NA	NA
		71	NA	96.7	NA	NA	NA
		72	NA	98.0	NA	NA	NA
		73	83.0	91.8	17	9	52.9
Reading	3	3	22	71.8	75.7	18	9
		3	67	65.6	58.6	19	7
Reading	4	4	22	71.5	56.2	48	23
		4	67	74.1	71.5	25	15
Reading	5	5	22	77.7	72.9	54	17
		5	67	77.5	77.6	26	14

Table 7-3: Montana Scoring Accuracy and Qualification Statistics 2007

Reading 6	6	22	78.0	62.2	34	24
	6	67	73.1	68.7	40	20
Reading 7	7	22	80.1	75.6	18	17
	7	67	81.1	76.5	23	21
Reading 8	8	22	75.0	71.2	22	11
	8	67	81.2	68.4	48	18
Reading 10	10	22	78.1	71.7	36	19
	10	72	71.4	78.1	41	16

TRAINING READERS

The QACs first applied the language of the scoring guide for an item to its anchor pack exemplars. Once discussion of the anchor pack had concluded, readers attempted to score the training pack exemplars correctly. The QACs then reviewed the training pack and answered any items readers had before actual scoring began. With this system, two aspects of scoring efficiency are in conflict. First, in order to minimize training expense, it is desirable to train each reader on as few items as possible. Second, to prevent reader drift and to minimize retraining requirements, it is desirable to score a given item in a brief period of time. But the lower the number of unique items each reader scores, the greater the number of readers required to score that item quickly. To minimize that conflict, we divided each subject area's readers into two or more groups. On the first day of scoring, each group was trained to score a different item. When a group had completed all of an item's responses, those readers were trained on another item (or set).

7.8 SCORING ACTIVITIES

Student test booklets at grade levels 3 through 8 and 10 were digitally scanned and scored on a file server for a dedicated, secure LAN. *iScore* then distributed digital images of student responses to readers. Training and scoring took place over a period of approximately two weeks.

Items were randomly assigned to readers; thus, each item in a student's response booklet was more than likely scored by a different reader. By using the maximum possible number of readers for each student, the procedure effectively minimized error variance due to reader sampling. All common

and matrix constructed-response items were scored once with a 2% read-behind to ensure consistency among readers and accuracy of individual readers.

Table 7-4: Montana 2007 Summary Statistics

Grade/Content	Number of Responses Scored	Total Number of Responses Scored in Double-Blind	Total Number of Arbitrations Required	Percentage of Double-Blinds Arbitrated
3 Math	83,922	2,575	80	3.11%
4 Math	82,885	4,336	270	6.23%
5 Math	84,090	3,355	159	4.74%
6 Math	85,662	3,709	219	5.90%
7 Math	90,250	5,101	237	4.65%
8 Math	91,621	5,250	237	4.51%
10 Math	93,058	8,059	355	4.41%
3 Reading	31,382	2,575	80	3.11%
4 Reading	31,322	4,336	270	6.23%
5 Reading	31,792	3,355	159	4.74%
6 Reading	32,442	3,709	219	5.90%
7 Reading	33,727	5,101	237	4.54%
8 Reading	34,655	5,250	237	5.69%
10 Reading	35,210	8,059	355	4.41%

7.9 MONITORING READERS

To ensure high inter-rater reliability and to prevent scoring drift after a reader scored a student response, *iScore* determined whether the reader met the accuracy requirement which is that a reader's scoring, based on double-scored responses, must be exact more than 90% of the time and that for the up to 10% that are not exact, their score is adjacent at least 80% of the time. If a reader's scores do not meet these three standards, *iScore* will freeze or block the reader's screen and alert the senior reader. The senior reader will then determine whether responses should also be scored by another reader, scored by a QAC, or routed for special attention. QAC's and senior readers were able to obtain current reader accuracy reports and speed reports online at any time. Table 7-4 summarizes how often a reader's screen was blocked through the process and the resolutions.

Table 7-5: Montana Blocked Reader Statistics 2007

Content	Grade/Item	Number of Readers Blocked From Scoring by *iScore*	Number of Readers Allowed to Continue Scoring Based upon Other Quality Monitoring (Read-Behinds and Double Blinds)	Number of Readers NOT Allowed To Continue Scoring Item and Reassigned to Other Items or Dismissed from Project
Math	3, 25	0	0	0
Math	3, 65	NA	NA	NA
Math	3, 66	NA	NA	NA
Math	3, 67	NA	NA	NA
Math	3, 68	0	0	0
Math	4, 25	0	0	0
Math	4, 65	NA	NA	NA
Math	4, 66	NA	NA	NA
Math	4, 67	NA	NA	NA
Math	4, 68	0	0	0
Math	5, 25	0	0	0
Math	5, 65	NA	NA	NA
Math	5, 66	NA	NA	NA
Math	5, 67	NA	NA	NA
Math	5, 68	0	0	0
Math	6, 25	1	1	0
Math	6, 65	NA	NA	NA
Math	6, 66	NA	NA	NA
Math	6, 67	NA	NA	NA
Math	6, 68	5	5	0
Math	7, 25	0	0	0
Math	7, 65	NA	NA	NA
Math	7, 66	NA	NA	NA
Math	7, 67	NA	NA	NA
Math	7, 68	0	0	0
Math	8, 25	0	0	0
Math	8, 65	NA	NA	NA
Math	8, 66	NA	NA	NA
Math	8, 67	NA	NA	NA
Math	8, 68	0	0	0
Math	10, 25	0	0	0
Math	10, 70	NA	NA	NA
Math	10,71	NA	NA	NA
Math	10, 72	NA	NA	NA
Math	10, 73	1	1	0
Reading	3, 22	1	1	0
Reading	3, 67	9	9	0
Reading	4, 22	9	8	1
Reading	4, 67	3	3	0
Reading	5, 22	5	5	0
Reading	5, 67	2	2	0
Reading	6, 22	4	4	0
Reading	6, 67	5	5	0

Table 7-5: Montana Blocked Reader Statistics 2007

Content	Grade/Item	Number of Readers Blocked From Scoring by *iScore*	Number of Readers Allowed to Continue Scoring Based upon Other Quality Monitoring (Read-Behinds and Double Blinds)	Number of Readers NOT Allowed To Continue Scoring Item and Reassigned to Other Items or Dismissed from Project
Reading	7, 22	2	2	0
Reading	7, 67	2	2	0
Reading	8, 22	2	2	0
Reading	8, 67	2	2	0
Reading	10, 22	2	2	0
Reading	10, 72	2	2	0
NOTE: All readers who were allowed to continue scoring did so under increased quality screening/additional read-behinds were conducted on these readers.				

7.10 GENERAL SCORING GUIDES

Tables 7-6 and 7-7 are examples of general CRT short-answer and constructed-response scoring guides.

Table 7-6: Short-Answer Items

Score Point	Description
1	The student's response provides a complete and correct answer.
0	The student's response is totally incorrect or too minimal to evaluate.
B	Blank/no response.

Table 7-7: Constructed- Response Items

Score Point	Description
4	• The student completes all important components of the task and communicates ideas clearly. • The student demonstrates in-depth understanding of the relevant concepts and/or processes. • When instructed to do so, the student chooses more efficient and/or sophisticated processes. • When instructed to do so, the student offers insightful interpretations or extensions (e.g., generalizations, applications, and analogies).
3	• The student completes the most important components of the task and communicates clearly. • The student demonstrates understanding of major concepts even though he/she overlooks or misunderstands some less important ideas or details.
2	• The student completes most important components of the task and communicates those clearly. • The student demonstrates that there are gaps in his/her conceptual understanding.
1	• The student shows minimal understanding. • The student addresses only a small portion of the required task(s).
0	• The student's response is totally incorrect or irrelevant.
B	• Blank/no response.

CHAPTER 8—ITEM ANALYSES

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each item. Both the *Standards for Educational and Psychological Testing (AERA et al., 1999)* and the *Code of Fair Testing Practices in Education (2004)* include standards for identifying quality items. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, items must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that Montana CRT items meet these standards. Qualitative analyses are described in earlier sections of this report; this section focuses on the more quantitative evaluations. The statistical evaluations are presented in three parts: 1) difficulty indices, 2) item-test correlations, and 3) differential item functioning (DIF) statistics. The item analyses presented here are based on the statewide administration of the Montana CRT in spring 2007. The numbers of students who participated in the assessment at each grade level were about 10,300 in grade 3, 10,200 in grade 4, 10,500 in grade 5, 10,550 in grade 6, 10,980 in grade 7, 11,130 in grade 8, and 11,170 in grade 10. Note that the information presented in this chapter is based on the items common to all forms since those are the items on which student scores are calculated. Item analyses are also performed for field test items; those statistics are then used in the item review process, as well as during form assembly for future administrations.

8.1 DIFFICULTY INDICES

All multiple-choice, constructed-response, and short-answer items were evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty was defined as the

average proportion of points achieved on an item, and was measured by obtaining the average score on an item and dividing by the maximum possible score for the item. Multiple-choice items were scored dichotomously (correct vs. incorrect), so for those items, the difficulty index is simply the proportion of students who correctly answered the item. Constructed-response items (two on each math form and two on each reading form) were scored polytomously, where a student can achieve a score of 0, 1, 2, 3, or 4. Short-answer items (three computation items on each math form) were scored 0 or 1. By computing the difficulty index as the average proportion of points achieved, the indices for the different item types are placed on a similar scale; the index ranges from 0.0 to 1.0 regardless of the item type. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an "easiness index" because larger values indicate easier items. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student ability, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students may indicate knowledge or skills that have not yet been mastered by most students, but such items provide little information about differences in student ability. In general, to provide best measurement, difficulty indices should range from near-chance performance (.25 for four-option, multiple-choice items or essentially zero for constructed-response or short-answer items) to .90. However, on a standards-referenced assessment such as the Montana CRT, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage (minimum of six items/points per standard).

8.2 ITEM DISCRIMINATION

A desirable feature of an item is that the higher-ability students perform better on the item than lower-ability students. The correlation between student performance on a single item and total test score is

a commonly used measure of this characteristic of an item. Within classical test theory, the item-test correlation is referred to as the item's discrimination because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For constructed-response items, the item discrimination index used was the Pearson product-moment correlation; for dichotomous items (multiple-choice and short-answer), the corresponding statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is –1.0 to +1.0, with a typical range from 0.2 to 0.6.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. Because each form of the Montana CRT was constructed to be parallel in content, the criterion score selected for each item was the raw score total for each form. The analyses were conducted for each form separately.

8.3 Summary of Item Analysis Results

Summary statistics of the difficulty and discrimination indices for each item are provided in Tables 8-1 through 8-7 for grades 3 through 8 and 10. Mean difficulty and discrimination indices, broken down by item type – multiple-choice, constructed-response (which includes both the four-point constructed-response and 1 one-point short-answer items), and all items – are shown in Table 8-8 (standard deviations are shown in parentheses). In general, the item difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There were a small number of items with near-zero discrimination indices, but none were reliably negative. While it is not

inappropriate to include items with low discrimination values or with very high or very low item difficulty values to ensure that content is appropriately covered, there were very few such cases on the Montana CRT.

A comparison of indices across grade levels is complicated because these indices are population dependent. Direct comparisons would require that either the items or students were common across groups. Since that is not the case, it can not be determined whether differences in performance across grade levels are due to differences in student ability or differences in item difficulty or both. However, one can say that for math, students in lower grades found their items somewhat less difficult than students in higher grades found their items.

Comparing the difficulty indices of multiple-choice items and constructed-response or short-answer items is inappropriate because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that the difficulty indices for multiple-choice items tend to be higher (indicating that students performed better on these items) than the difficulty indices for constructed-response items. Similarly, the partial credit allowed by four-point constructed-response items is advantageous in the computation of item-test correlations, so the discrimination indices for these items tend to be larger than the discrimination indices of multiple-choice or short-answer items.

The statistics in Tables 8-1 through 8-7 and those calculated for the full set of items in Table 8-8 are weighted according to the number of points contributed by each item. In the event that an item's statistics indicate it is flawed, the item is dropped from the scoring of the operational form. An item may be dropped, for example, if more than one of the response options is a defensible answer, or if the item is misleading or unclear in some way. One flawed item was found for the 2007 MontCAS, Phase 2 CRT test administration in Grade 7 mathematics.

Table 8-1: Item Analysis: Grade 3

Content Area		Difficulty	Discrimination
Reading	Mean	0.67	0.37
	StDev	0.13	0.08
	Min	0.35	0.15
	Max	0.91	0.52
	Range	0.56	0.37
Math	Mean	0.69	0.35
	StDev	0.16	0.10
	Min	0.23	0.07
	Max	0.93	0.54
	Range	0.70	0.47

Table 8-2: Item Analysis: Grade 4

Content Area		Difficulty	Discrimination
Reading	Mean	0.67	0.36
	StDev	0.13	0.09
	Min	0.40	0.15
	Max	0.93	0.54
	Range	0.53	0.39
Math	Mean	0.63	0.36
	StDev	0.13	0.08
	Min	0.29	0.22
	Max	0.88	0.52
	Range	0.59	0.3

Table 8-3: Item Analysis: Grade 5

Content Area		Difficulty	Discrimination
Reading	Mean	0.70	0.36
	StDev	0.14	0.09
	Min	0.36	0.08
	Max	0.94	0.53
	Range	0.58	0.45
Math	Mean	0.60	0.37
	StDev	0.15	0.09
	Min	0.21	0.20
	Max	0.86	0.62
	Range	0.65	0.42

Table 8-4: Item Analysis: Grade 6

Content Area		Difficulty	Discrimination
Reading	Mean	0.70	0.34
	StDev	0.14	0.08
	Min	0.35	0.05
	Max	0.95	0.49
	Range	0.60	0.44
Math	Mean	0.58	0.35
	StDev	0.16	0.09
	Min	0.20	0.17
	Max	0.92	0.53
	Range	0.72	0.36

Table 8-5: Item Analysis: Grade 7

Content Area		Difficulty	Discrimination
Reading	Mean	0.70	0.36
	StDev	0.11	0.08
	Min	0.37	0.15
	Max	0.88	0.49
	Range	0.51	0.34
Math	Mean	0.54	0.34
	StDev	0.16	0.09
	Min	0.22	0.18
	Max	0.90	0.63
	Range	0.68	0.45

Table 8-6: Item Analysis: Grade 8

Content Area		Difficulty	Discrimination
Reading	Mean	0.72	0.38
	StDev	0.11	0.08
	Min	0.42	0.17
	Max	0.93	0.54
	Range	0.51	0.37
Math	Mean	0.56	0.40
	StDev	0.14	0.10
	Min	0.21	0.21
	Max	0.88	0.72
	Range	0.67	0.51

Table 8-7: Item Analysis: Grade 10

Content Area		Difficulty	Discrimination
Reading	Mean	0.71	0.36
	StDev	0.14	0.09
	Min	0.42	0.15
	Max	0.95	0.50
	Range	0.53	0.35
Math	Mean	0.52	0.36
	StDev	0.17	0.09
	Min	0.21	0.20
	Max	0.89	0.65
	Range	0.68	0.45

Table 8-8: Average Difficulty and Discrimination of Different Item Types For Each Grade/Content Area Combination

Grade	Content Area		Item Type		
			All	MC	Constructed-Response
3	Reading	Difficulty	0.67 (0.13)	0.68 (0.12)	0.39 (0.05)
		Discrimination	0.37 (0.08)	0.37 (0.08)	0.43 (0.07)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.69 (0.16)	0.69 (0.17)	0.70 (0.16)
		Discrimination	0.35 (0.10)	0.34 (0.10)	0.43 (0.08)
		Number of Items	60	55	5
4	Reading	Difficulty	0.67 (0.13)	0.68 (0.13)	0.43 (0.04)
		Discrimination	0.36 (0.09)	0.36 (0.09)	0.41 (0.08)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.63 (0.13)	0.64 (0.13)	0.52 (0.12)
		Discrimination	0.36 (0.08)	0.36 (0.07)	0.44 (0.06)
		Number of Items	60	55	5
5	Reading	Difficulty	0.70 (0.14)	0.71 (0.13)	0.45 (0.01)
		Discrimination	0.36 (0.09)	0.36 (0.09)	0.45 (0.11)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.60 (0.15)	0.61 (0.14)	0.51 (0.24)
		Discrimination	0.37 (0.09)	0.36 (0.08)	0.45 (0.14)
		Number of Items	60	55	5
6	Reading	Difficulty	0.70 (0.14)	0.71 (0.14)	0.51 (0.09)
		Discrimination	0.34 (0.08)	0.34 (0.08)	0.44 (0.06)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.58 (0.16)	0.59 (0.16)	0.47 (0.13)
		Discrimination	0.35 (0.09)	0.34 (0.09)	0.43 (0.08)
		Number of Items	60	55	5
7	Reading	Difficulty	0.70 (0.11)	0.70 (0.11)	0.57 (0.03)
		Discrimination	0.36 (0.08)	0.35 (0.08)	0.48 (0.02)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.54 (0.16)	0.56 (0.16)	0.38 (0.12)
		Discrimination	0.34 (0.09)	0.33 (0.08)	0.49 (0.09)
		Number of Items	59	54	5
8	Reading	Difficulty	0.72 (0.11)	0.73 (0.10)	0.44 (0.02)
		Discrimination	0.38 (0.08)	0.38 (0.07)	0.52 (0.04)
		Number of Items	54	52	2
	Mathematics	Difficulty	0.56 (0.14)	0.56 (0.14)	0.50 (0.12)
		Discrimination	0.40 (0.10)	0.38 (0.08)	0.57 (0.13)
		Number of Items	60	55	5
10	Reading	Difficulty	0.71 (0.14)	0.72 (0.13)	0.49 (0.03)
		Discrimination	0.36 (0.09)	0.35 (0.08)	0.49 (0.00)
		N	59	57	2
	Mathematics	Difficulty	0.52 (0.17)	0.53 (0.17)	0.38 (0.17)
		Discrimination	0.36 (0.09)	0.35 (0.08)	0.52 (0.09)
		N	65	60	5

Note: Numbers shown in parentheses are standard deviations

8.4 DIFFERENTIAL ITEM FUNCTIONING (DIF)

The *Code of Fair Testing Practices in Education* (2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit, and actions should be taken

to make certain that differences in performance are due to construct-relevant, rather than irrelevant, factors. The *Standards for Educational and Psychological Testing* (AERA et al., 1999) includes similar guidelines. As part of the effort to identify such problems, Montana CRT items were evaluated in terms of differential item functioning (DIF) statistics.

DIF procedures are designed to identify items for which subgroups of interest perform differently beyond the impact of differences in overall achievement. For the Montana CRT, the standardization DIF procedure (Dorans and Kulick, 1986) was employed to evaluate subgroup differences for three comparison groups: male/female, white/Native American, and white/Hispanic. This procedure calculates the difference in item performance for groups of students matched for achievement on the total test. That is, the average item performance is calculated for students at every total score, then an overall average is calculated weighting by the total score distribution so the weighting is the same for the two groups. The index ranges from –1.00 to 1.00 for multiple-choice and short-answer items and is adjusted to the same scale for constructed-response items. Negative numbers indicate that the item was more difficult for female or non-white students. Dorans and Holland (1993) suggested that index values between –0.05 and 0.05 should be considered negligible. Most Montana CRT items fall within this range. Dorans and Holland further stated that items with values between –0.10 and –0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked, and that items with values outside the [–0.10, 0.10] range (i.e., "high" DIF) are more unusual and should be examined very carefully.

DIF indices indicate the degree of differential performance between two groups. That differential performance may or may not be indicative of bias in the test. Course-taking patterns, group differences in interests, or differences in school curricula can lead to DIF. If subgroup differences in performance are related to construct-relevant factors, the items should be considered for inclusion on a test.

Each item was categorized according to the guidelines adapted from Dorans and Holland (1993). Table 8-9 shows the number of items classified into each category separately by item type (multiple-choice versus constructed-response; open-response items are included with constructed-response). Results are shown for male/female, white/Native American, and white/Hispanic comparisons. Table 8-10 provides the number of items in each of the three DIF categories that favor males or females, also separately by item type (multiple-choice and constructed-response; open-response items are included with constructed-response). There are some Montana CRT items categorized as "low" or "high" DIF. These indices must not be interpreted as indisputable evidence of bias. Both the *Code of Fair Testing Practices in Education* (2004) and the *Standards for Educational and Psychological Testing* (AERA et al., 1999) assert that test items must be free from construct-irrelevant sources of differential difficulty. If subgroup differences in performance can be plausibly attributed to construct-relevant factors, the items may be included on a test. What is important is to determine if the cause of this differential performance is construct-relevant.

For the Montana CRT, there were relatively few items (less than five) flagged as having low or high DIF. The items that were flagged were reviewed for potential bias, and no obvious biases were detected. For this reason, and in order to ensure sufficient content coverage, no items were excluded from the test as a result of the DIF analyses.

Table 8-9: DIF Analysis: All Grades

Grade	Content Area	Male/Female DIF Class									White/Native American DIF Class									White/Hispanic DIF Class								
		All			MC			CR			All			MC			CR			All			MC			CR		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
3	Reading	45	9	0	43	9	0	2	0	0	49	5	0	47	5	0	2	0	0	50	4	0	48	4	0	2	0	0
	Math	55	5	0	50	5	0	5	0	0	56	4	0	51	4	0	5	0	0	54	4	2	49	4	2	5	0	0
4	Reading	46	5	3	45	4	3	1	1	0	50	4	0	48	4	0	2	0	0	49	5	0	47	5	0	2	0	0
	Math	54	5	1	49	5	1	5	0	0	56	4	0	52	3	0	4	1	0	48	12	0	44	11	0	4	1	0
5	Reading	49	5	0	48	4	0	1	1	0	51	3	0	49	3	0	2	0	0	52	2	0	50	2	0	2	0	0
	Math	53	7	0	48	7	0	5	0	0	57	3	0	52	3	0	5	0	0	58	2	0	53	2	0	5	0	0
6	Reading	47	5	2	46	4	2	1	1	0	49	5	0	47	5	0	2	0	0	49	5	0	47	5	0	2	0	0
	Math	50	8	2	47	6	2	3	2	0	57	3	0	54	1	0	3	2	0	55	5	0	51	4	0	4	1	0
7	Reading	42	8	4	42	6	4	0	2	0	47	7	0	45	7	0	2	0	0	52	2	0	50	2	0	2	0	0
	Math	49	9	1	45	8	1	4	1	0	56	3	0	51	3	0	5	0	0	52	7	0	47	7	0	5	0	0
8	Reading	45	7	2	45	5	2	0	2	0	50	4	0	48	4	0	2	0	0	48	6	0	46	6	0	2	0	0
	Math	49	10	1	45	9	1	4	1	0	55	5	0	51	4	0	4	1	0	50	10	0	45	10	0	5	0	0
10	Reading	50	9	0	50	7	0	0	2	0	51	7	1	49	7	1	2	0	0	55	4	0	53	4	0	2	0	0
	Math	49	13	3	45	12	3	4	1	0	60	5	0	55	5	0	5	0	0	57	8	0	53	7	0	4	1	0

A = negligible DIF, B = low DIF, C = high DIF

TABLE 8-10: MALE VS. FEMALE DIFFERENTIAL ITEM FUNCTIONING (DIF) CATEGORIZATION BY ITEM TYPE (MULTIPLE-CHOICE AND CONSTRUCTED-RESPONSE)[1]

			Negligible DIF (A)				Low DIF (B)				High DIF (C)			
Grade	Content Area	Item Type	Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%
3	Reading	MC	30	13	43	83	4	5	9	17	0	0	0	0
		CR	2	0	2	100	0	0	0	0	0	0	0	0
	Math	MC	25	25	50	91	0	5	5	9	0	0	0	0
		CR	4	1	5	100	0	0	0	0	0	0	0	0
4	Reading	MC	30	15	45	87	1	3	4	8	0	3	3	6
		CR	1	0	1	50	1	0	1	50	0	0	0	0
	Math	MC	27	22	49	89	2	3	5	9	0	1	1	2
		CR	4	1	5	100	0	0	0	0	0	0	0	0
5	Reading	MC	30	18	48	92	1	3	4	8	0	0	0	0
		CR	1	0	1	50	1	0	1	50	0	0	0	0
	Math	MC	29	19	48	87	3	4	7	13	0	0	0	0
		CR	4	1	5	100	0	0	0	0	0	0	0	0
6	Reading	MC	26	20	46	88	2	2	4	8	0	2	2	4
		CR	1	0	1	50	1	0	1	50	0	0	0	0
	Math	MC	23	24	47	85	1	5	6	11	0	2	2	4
		CR	3	0	3	60	2	0	2	40	0	0	0	0
7	Reading	MC	24	18	42	81	2	4	6	12	1	3	4	8
		CR	0	0	0	0	2	0	2	100	0	0	0	0
	Math	MC	27	18	45	83	2	6	8	15	0	1	1	2
		CR	3	1	4	80	1	0	1	20	0	0	0	0
8	Reading	MC	28	17	45	87	2	3	5	10	0	2	2	4
		CR	0	0	0	0	2	0	2	100	0	0	0	0
	Math	MC	22	23	45	82	4	5	9	16	0	1	1	2
		CR	4	0	4	80	1	0	1	20	0	0	0	0
10	Reading	MC	32	18	50	88	1	6	7	12	0	0	0	0
		CR	0	0	0	0	2	0	2	100	0	0	0	0
	Math	MC	29	16	45	75	4	8	12	20	1	2	3	5
		CR	3	1	4	80	1	0	1	20	0	0	0	0

8.5 DIMENSIONALITY ANALYSES

The DIF analyses of the previous section were performed to identify items which showed evidence of differences in performance between pairs of subgroups beyond that which would be expected based on the primary construct that underlies total test score (also known as the "primary dimension;" for example, general achievement in math). When items are flagged for DIF, statistical evidence points to their measuring an additional dimension(s) to the primary dimension.

Because tests are constructed with multiple content area subcategories, and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the

[1] The percents reported in Table 8-10 are percents of total items.

common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional IRT models that are used for calibrating, linking, scaling, and equating the 2006-07 MontCAS test forms. As noted in the previous section, a statistically significant DIF result does not automatically imply that an item is measuring an *irrelevant* construct or dimension. An item could be flagged for DIF because it measures one of the construct-*relevant* dimensions of a subcategory's knowledge and skills.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Findings from dimensionality (DIM) analyses performed on the 2006-07 MontCAS common items for Math and Reading are reported below. (Note: only common items were analyzed since they are used for score reporting.)

The DIM analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both of these methods use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on expected total score for the rest of the test, and the average conditional covariance is obtained by averaging over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected total test scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and local *dependence* implies multidimensionality. Thus, non-random patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first divided into a training sample and a cross-validation sample. Then an exploratory analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items displays local dependence, conditioning on total score on the non-clustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first divided into a training sample and a cross-validation sample. (Note: The random training and cross-validation samples used for the DIMTEST analyses were drawn independently of the sample used for the DETECT analyses.) The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed, from this sum the between-cluster conditional covariances are subtracted, this difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality), values of 0.2 to 0.4 weak to moderate multidimensionality; values of 0.4 to 1.0 moderate to high multidimensionality, and values greater than 1.0 strong multidimensionality.

DIMTEST and DETECT were applied to the 2006-07 MontCAS. The data for each grade and content area were split into a training sample and a cross-validation sample. Every grade/content area combination had at least 10,000 student examinees, so every training sample and cross-validation

sample had at least 5000 students. DIMTEST was then applied to every grade/content area. DETECT was applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

Because of the large sample sizes of the Montana tests, DIMTEST would be sensitive even to quite small violations of unidimensionality, and the null hypothesis was strongly rejected for every dataset ($p \leq 0.00005$ for every grade/content area). These results were not surprising because strict unidimensionality is an idealization that almost never holds exactly for a given dataset. Thus, it was important to use DETECT to estimate the effect size of the violations of local independence found by DIMTEST. Table 8-11 displays the multidimensional effect size estimates from DETECT.

Table 8-11. 2006-07 MontCAS: Multidimensionality Effect Sizes by Grade and Subject.

Grade	Subject	Multidimensionality Effect Size
3	Math	0.13
	Reading	0.10
4	Math	0.12
	Reading	0.10
5	Math	0.16
	Reading	0.11
6	Math	0.15
	Reading	0.10
7	Math	0.14
	Reading	0.14
8	Math	0.11
	Reading	0.11
10	Math	0.14
	Reading	0.11

All the DETECT values indicated very weak multidimensionality. The Math test forms (average effect size of about 0.14) tended to show slightly greater multidimensionality than did Reading (average of about 0.11). Such small violations of local independence do not warrant any changes in test design or scoring.

8.6 ITEM RESPONSE THEORY ANALYSES

In addition to the classical test theory item analyses previously described, the Montana CRT tests were analyzed according to item response theory (IRT) models. IRT analyses were used, first, to place all 2007 forms on the same scale, and second, to equate the 2007 test to the previous year's test. Details on the IRT calibration and equating procedures for the Montana CRT are provided in Chapter 10.

CHAPTER 9—RELIABILITY

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together and complement one another. Tests that function well provide a dependable assessment of the student's level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either higher or lower than his or her true ability. For example, a student may mis-read an item, or mistakenly fill in the wrong bubble when he or she knew the answer; similarly a student may get an item correct by guessing, even though he or she did not know the answer. Collectively, these extraneous factors that impact a student's score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores or vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors made are small on average and student scores on such a test will consistently represent their ability) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable. (This is referred to as test-retest reliability.) A potential problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the 'remembering items' problem is to give a different, but parallel test at the second administration. If student scores on each test correlate highly

the test is considered reliable. (This is known as alternate forms reliability, because an alternate form of the test is used in each administration.) This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel forms reliability indices. One way to address these problems is to split the test in half and then correlate students' scores on the two half-tests; this in effect treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval, and of creating and administering two parallel forms of the test, are alleviated. This is known as a split-half estimate of reliability. If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires a judgment regarding the selection of which items contribute to which half-test score. This decision may have an impact on the resulting correlation; different splits will give different estimates of reliability. Cronbach (1951) provided a statistic,α, which avoids this concern about the split-half method. Cronbach's α gives an estimate of the average of all possible splits for a given test. Cronbach's α is often referred to as a measure of internal consistency because it provides a measure of how well the items in a test are intercorrelated. Cronbach's α is computed using the following formula:

$$\alpha \equiv \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{n}\sigma^2_{(Y_i)}}{\sigma^2_x}\right]$$

where i indexes the item
n is the total number of items,
$\sigma^2_{(Y_i)}$ represents individual item variance, and
σ^2_x represents the total test variance

In addition to Cronbach's α, another approach to estimating the reliability for a test with differing item types (i.e., multiple-choice and constructed-response) is to assume that at least a small, but important, degree of unique variance is associated with item type (Feldt and Brennan, 1989). In contrast, Cronbach's coefficient α is built upon the assumption that there are no such local or clustered dependencies. A stratified version of coefficient α corrects for this problem by using the following formula:

$$\alpha_{strat} = 1 - \frac{\sum_{j=1}^{k} \sigma_{x_j}^2 (1 - \alpha_j)}{\sigma_x^2}$$

where j indexes the subtests or categories,
$\sigma_{x_j}^2$ represents the variance of each of the k individual subtests or categories,
α_j is the unstratified Cronbach's α coefficient for each subtest, and
σ_x^2 represents the total test variance.

9.1 Reliability and Standard Errors of Measurement

Table 9-1 provides descriptive statistics, the overall Cronbach's α coefficient for each grade/content combination, and raw score standard errors of measurement. Tables 9-2 through 9-8 present Cronbach's α for each test form in each subject area (reading and mathematics), separately for each grade level. The tables also show reliability coefficients separately for multiple-choice and constructed-response (which includes short-answer in mathematics) items, and stratified reliability coefficients that adjust for the fact that different item formats are included in the test.

Across the grades and content areas, the overall α coefficients, multiple-choice α coefficients, and stratified α coefficients range from the upper-.80s to the low-.90s. There are little or no differences between the overall α and stratified α coefficients. The α coefficients for the constructed-response items are substantially lower, ranging from around 0.40 to around 0.70. These lower values can be explained, at least to some extent, by the fact that there are greater scoring inconsistencies for

constructed-response items, as well as the relatively small numbers of these items on the test. Note that, for reading, it is possible that the reliability coefficients are inflated as a result of passage-based item dependency.

Table 9-1: Reliabilities, Standard Errors of Measurement, and Descriptive Statistics

Grade	Content Area	N	Total Points	Mean	SD	Rel	SEM
3	Reading	10259	60	45.52	10.76	0.89	3.50
	Mathematics	10303	66	38.61	10.26	0.90	3.19
4	Reading	10168	60	40.66	11.88	0.91	3.62
	Mathematics	10204	66	38.52	9.97	0.89	3.25
5	Reading	10506	60	39.24	12.26	0.91	3.69
	Mathematics	10528	66	40.57	9.86	0.90	3.18
6	Reading	10554	60	37.19	11.86	0.90	3.76
	Mathematics	10570	66	40.81	9.30	0.89	3.15
7	Reading	10975	60	34.73	11.51	0.89	3.77
	Mathematics	10979	65	40.92	10.04	0.90	3.23
8	Reading	11133	60	35.99	13.45	0.92	3.72
	Mathematics	11127	66	41.47	10.30	0.91	3.15
10	Reading	11174	65	35.75	13.19	0.91	3.88
	Mathematics	11164	71	45.00	10.38	0.90	3.28

Table 9-2: Reliability Analysis – Grade 3

Content Area	Reliability	Form															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reading	Coeff α	0.91	0.91	0.90	0.90	0.91	0.90	0.90	0.90	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.90
	MC α	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.91	0.90	0.90	0.89	0.90	0.89	0.90
	CR α	0.52	0.52	0.47	0.42	0.50	0.53	0.45	0.49	0.52	0.48	0.52	0.49	0.38	0.44	0.41	0.40
	Strat α	0.91	0.91	0.90	0.90	0.91	0.90	0.90	0.90	0.91	0.91	0.91	0.91	0.90	0.90	0.90	0.90
Mathe-matics	Coeff α	0.90	0.90	0.89	0.90	0.90	0.89	0.89	0.88	0.90	0.90	0.89	0.90	0.89	0.89	0.88	0.89
	MC α	0.90	0.90	0.88	0.89	0.90	0.89	0.89	0.87	0.89	0.89	0.88	0.89	0.89	0.88	0.88	0.89
	CR α	0.52	0.51	0.49	0.49	0.53	0.49	0.48	0.49	0.50	0.51	0.51	0.51	0.48	0.53	0.47	0.49
	Strat α	0.90	0.90	0.89	0.90	0.90	0.89	0.90	0.88	0.90	0.90	0.89	0.90	0.89	0.90	0.88	0.90

Table 9-3: Reliability Analysis – Grade 4

Content Area	Reliability	Form															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reading	Coeff α	0.89	0.89	0.90	0.90	0.89	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.89	0.89	0.90	0.89
	MC α	0.89	0.89	0.89	0.90	0.89	0.89	0.89	0.88	0.89	0.89	0.90	0.89	0.89	0.89	0.89	0.88
	CR α	0.56	0.55	0.58	0.57	0.52	0.52	0.57	0.52	0.53	0.56	0.51	0.60	0.62	0.56	0.55	0.54
	Strat α	0.90	0.90	0.90	0.90	0.89	0.90	0.89	0.89	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.89
Mathe-matics	Coeff α	0.90	0.91	0.92	0.91	0.91	0.91	0.91	0.90	0.91	0.91	0.91	0.90	0.90	0.91	0.91	0.90
	MC α	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.89
	CR α	0.57	0.55	0.59	0.60	0.55	0.58	0.57	0.58	0.60	0.59	0.58	0.55	0.54	0.55	0.56	0.59
	Strat α	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.90	0.91	0.91	0.91	0.90

Table 9-4: Reliability Analysis – Grade 5

Content Area	Reliability	Form															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reading	Coeff α	0.90	0.91	0.89	0.89	0.91	0.90	0.89	0.90	0.89	0.89	0.90	0.88	0.90	0.90	0.90	0.89
	MC α	0.90	0.90	0.89	0.88	0.91	0.90	0.89	0.89	0.89	0.89	0.90	0.88	0.90	0.89	0.90	0.89
	CR α	0.48	0.54	0.50	0.46	0.49	0.49	0.40	0.47	0.40	0.44	0.46	0.45	0.49	0.49	0.45	0.45
	Strat α	0.90	0.91	0.90	0.89	0.91	0.90	0.89	0.90	0.89	0.89	0.90	0.89	0.90	0.90	0.90	0.90
Mathe-matics	Coeff α	0.91	0.90	0.91	0.90	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.90	0.91
	MC α	0.91	0.90	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.90
	CR α	0.55	0.52	0.56	0.56	0.58	0.58	0.56	0.59	0.54	0.56	0.54	0.52	0.55	0.56	0.53	0.58
	Strat α	0.92	0.90	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.90	0.91

Table 9-5: Reliability Analysis – Grade 6

Content Area	Reliability	Form															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reading	Coeff α	0.90	0.89	0.89	0.88	0.88	0.89	0.88	0.87	0.88	0.89	0.88	0.88	0.88	0.88	0.90	0.88
	MC α	0.89	0.89	0.88	0.88	0.88	0.88	0.88	0.86	0.87	0.88	0.87	0.88	0.87	0.88	0.89	0.88
	CR α	0.60	0.57	0.54	0.50	0.57	0.54	0.60	0.55	0.58	0.51	0.57	0.55	0.53	0.54	0.65	0.57
	Strat α	0.90	0.90	0.89	0.89	0.89	0.89	0.89	0.88	0.88	0.89	0.88	0.89	0.88	0.89	0.90	0.89
Mathe-matics	Coeff α	0.90	0.90	0.89	0.89	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.89	0.90	0.91	0.90
	MC α	0.89	0.90	0.89	0.89	0.89	0.89	0.90	0.89	0.90	0.89	0.89	0.89	0.88	0.89	0.90	0.89
	CR α	0.55	0.54	0.52	0.53	0.51	0.53	0.53	0.50	0.52	0.52	0.53	0.53	0.51	0.51	0.55	0.49
	Strat α	0.91	0.91	0.89	0.90	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.89	0.90	0.91	0.90

Table 9-6: Reliability Analysis – Grade 7

Content Area	Reliability	Form															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reading	Coeff α	0.90	0.89	0.90	0.90	0.89	0.90	0.91	0.90	0.89	0.90	0.90	0.89	0.89	0.89	0.90	0.90
	MC α	0.90	0.89	0.90	0.89	0.88	0.89	0.90	0.89	0.89	0.89	0.89	0.88	0.89	0.89	0.90	0.89
	CR α	0.69	0.63	0.69	0.64	0.59	0.68	0.70	0.69	0.71	0.65	0.62	0.68	0.66	0.68	0.72	0.66
	Strat α	0.91	0.90	0.91	0.90	0.89	0.90	0.91	0.91	0.90	0.90	0.90	0.89	0.90	0.90	0.91	0.90
Mathe-matics	Coeff α	0.90	0.89	0.89	0.89	0.89	0.90	0.89	0.89	0.89	0.88	0.89	0.88	0.89	0.89	0.89	0.90
	MC α	0.89	0.88	0.88	0.87	0.87	0.89	0.87	0.88	0.88	0.87	0.88	0.87	0.88	0.88	0.88	0.88
	CR α	0.62	0.59	0.60	0.55	0.58	0.56	0.59	0.57	0.55	0.55	0.59	0.54	0.58	0.55	0.57	0.59
	Strat α	0.90	0.90	0.90	0.89	0.89	0.90	0.89	0.89	0.89	0.89	0.89	0.89	0.90	0.89	0.89	0.90

Table 9-7: Reliability Analysis – Grade 8

Content Area	Reliability	Form															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reading	Coeff α	0.92	0.91	0.91	0.90	0.90	0.91	0.90	0.92	0.90	0.91	0.91	0.90	0.90	0.90	0.90	0.90
	MC α	0.92	0.91	0.90	0.90	0.90	0.91	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.89
	CR α	0.71	0.66	0.67	0.68	0.68	0.68	0.67	0.69	0.67	0.66	0.68	0.71	0.69	0.67	0.68	0.65
	Strat α	0.92	0.92	0.91	0.91	0.91	0.92	0.91	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.90
Mathe-matics	Coeff α	0.93	0.92	0.92	0.93	0.93	0.92	0.92	0.93	0.92	0.93	0.92	0.92	0.92	0.92	0.92	0.92
	MC α	0.91	0.91	0.91	0.92	0.92	0.91	0.91	0.92	0.91	0.92	0.91	0.91	0.91	0.91	0.91	0.91
	CR α	0.71	0.69	0.65	0.65	0.67	0.66	0.66	0.68	0.67	0.68	0.64	0.63	0.67	0.66	0.67	0.65
	Strat α	0.93	0.93	0.92	0.93	0.93	0.93	0.92	0.93	0.92	0.93	0.92	0.92	0.93	0.92	0.92	0.92

Table 9-8: Reliability Analysis – Grade 10

Content Area	Reliability	Form															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Reading	Coeff α	0.92	0.90	0.91	0.91	0.91	0.89	0.89	0.89	0.89	0.89	0.88	0.90	0.91	0.90	0.90	0.90
	MC α	0.91	0.89	0.90	0.91	0.90	0.89	0.89	0.89	0.89	0.89	0.87	0.90	0.90	0.90	0.90	0.90
	CR α	0.73	0.72	0.69	0.70	0.71	0.66	0.71	0.72	0.64	0.71	0.70	0.65	0.71	0.66	0.67	0.69
	Strat α	0.92	0.90	0.91	0.92	0.91	0.90	0.90	0.90	0.90	0.90	0.89	0.91	0.91	0.91	0.91	0.91
Mathe-matics	Coeff α	0.92	0.91	0.91	0.92	0.92	0.92	0.92	0.91	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.92
	MC α	0.90	0.90	0.90	0.91	0.91	0.90	0.90	0.90	0.89	0.90	0.90	0.91	0.90	0.90	0.90	0.91
	CR α	0.64	0.58	0.61	0.63	0.63	0.60	0.62	0.56	0.60	0.57	0.60	0.61	0.61	0.61	0.60	0.60
	Strat α	0.92	0.91	0.91	0.92	0.92	0.92	0.92	0.91	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.92

9.2 SUBGROUP RELIABILITY

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2006-07 Montana CRT assessments. Appendix F presents reliabilities for various subgroups of interest. Subgroup Cronbach's α's were calculated using the

formula defined above including only the members of the subgroup in question in the computations. For mathematics, subgroup reliabilities ranged from 0.78 to 0.92, and for reading from 0.81 to 0.91.

For several reasons, the results of this subsection should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but on the statistical distribution of the studied subgroup. For example, subgroup sample sizes may vary considerably (see Appendix F), resulting in natural variation in reliability coefficients. Alpha, a type of correlation coefficient, may be artificially depressed for subgroups with little variability (Draper & Smith, 1998). Finally, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup.

9.3 Reporting Subcategories Reliability

In previous sections, the reliability coefficients were calculated based on form and item type. Item type represents just one way of breaking an overall test into subtests. Of even more interest are reliabilities for the reporting subcategories within Montana CRT subject areas, described in Chapters 4 and 5. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Table 9-9. Once again as expected, because they are based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this information into account.

Table 9-9: 2006-07 Montana CRT Common Item α by Grade, Subject, and Reporting Subcategory.

Grade	Subject	Reporting Subcategory	Possible Points	α
3	Math	Problem Solving	8	0.51
		Numbers and Operations	13	0.69
		Algebra	6	0.46
		Geometry	11	0.52
		Measurement	9	0.56
		Data Analysis, Statistics, and Probability	12	0.52
		Patterns, Relations, and Functions	7	0.65
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	23	0.84
		Students apply a range of skills and strategies to read	19	0.68
		Students select, read and respond to print and non-print material for a variety of purposes	10	0.60
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	8	0.47
4	Math	Problem Solving	9	0.57
		Numbers and Operations	12	0.71
		Algebra	6	0.51
		Geometry	10	0.58
		Measurement	8	0.58
		Data Analysis, Statistics, and Probability	13	0.57
		Patterns, Relations, and Functions	8	0.61
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	19	0.74
		Students apply a range of skills and strategies to read	20	0.76
		Students select, read and respond to print and non-print material for a variety of purposes	8	0.54
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	13	0.58
5	Math	Problem Solving	8	0.52
		Numbers and Operations	14	0.74
		Algebra	6	0.55
		Geometry	10	0.56
		Measurement	9	0.46
		Data Analysis, Statistics, and Probability	13	0.70
		Patterns, Relations, and Functions	6	0.57
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	18	0.69
		Students apply a range of skills and strategies to read	24	0.77
		Students select, read and respond to print and non-print material for a variety of purposes	8	0.51
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	10	0.71

Table 9-10. 2006-07 Montana CRT Common Item α by Grade, Subject, and Reporting Subcategory (cont'd).

Grade	Subject	Reporting Subcategory	Possible Points	α
6	Math	Problem Solving	8	0.46
		Numbers and Operations	14	0.71
		Algebra	7	0.59
		Geometry	9	0.49
		Measurement	9	0.57
		Data Analysis, Statistics, and Probability	13	0.59
		Patterns, Relations, and Functions	6	0.55
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	19	0.70
		Students apply a range of skills and strategies to read	22	0.72
		Students select, read and respond to print and non-print material for a variety of purposes	8	0.56
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	11	0.60
7	Math	Problem Solving	7	0.47
		Numbers and Operations	10	0.67
		Algebra	8	0.57
		Geometry	11	0.50
		Measurement	6	0.40
		Data Analysis, Statistics, and Probability	13	0.58
		Patterns, Relations, and Functions	10	0.61
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	22	0.74
		Students apply a range of skills and strategies to read	18	0.69
		Students select, read and respond to print and non-print material for a variety of purposes	10	0.62
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	10	0.66
8	Math	Problem Solving	8	0.47
		Numbers and Operations	10	0.72
		Algebra	8	0.73
		Geometry	13	0.69
		Measurement	7	0.55
		Data Analysis, Statistics, and Probability	13	0.64
		Patterns, Relations, and Functions	7	0.61
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	20	0.72
		Students apply a range of skills and strategies to read	19	0.78
		Students select, read and respond to print and non-print material for a variety of purposes	9	0.64
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	12	0.70

Table 9-10. 2006-07 Montana CRT Common Item α by Grade, Subject, and Reporting Subcategory

Grade	Subject	Reporting Subcategory	Possible Points	α
10	Math	Problem Solving	7	0.47
		Numbers and Operations	10	0.56
		Algebra	10	0.73
		Geometry	13	0.67
		Measurement	8	0.49
		Data Analysis, Statistics, and Probability	13	0.63
		Patterns, Relations, and Functions	10	0.63
	Reading	Students construct meaning as they comprehend, interpret, and respond to what they read	18	0.73
		Students apply a range of skills and strategies to read	22	0.76
		Students select, read and respond to print and non-print material for a variety of purposes	13	0.62
		Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	12	0.62

For mathematics, subcategory reliabilities ranged from 0.40 to 0.74, and for reading from 0.47 to 0.84. In general, the subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect based on classical test theory. Qualitative differences between grades and content areas once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

9.4 Reliability of Performance Level Categorization

All test scores contain measurement error; thus classifications based on test scores are also subject to measurement error. After the performance levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the Montana CRT, students are classified into one of four performance levels: *Novice* (N), *Nearing Proficiency* (NP), *Proficient* (P), or *Advanced* (A). This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

9.5 ACCURACY

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist.

9.6 CONSISTENCY

Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete, parallel forms of the test are given to the same group of students. This is usually impractical, especially on lengthy tests. To overcome this issue, techniques have been developed to estimate both accuracy and consistency of classification decisions based on a single administration of a test. The technique developed by Livingston and Lewis (1995) was used for the Montana CRT because their technique can be used with both constructed-response and multiple-choice items.

9.7 CALCULATING ACCURACY

All of the accuracy and consistency estimation techniques described below make use of the concept of "true scores" in the sense of classical test theory. A true score is the score that would be obtained on a test that had no measurement error. It is a theoretical concept that cannot be observed, although it can be estimated. In the Livingston and Lewis method, the estimated true score distribution is used to estimate the proportion of students in each "true" performance level. After various technical adjustments (which are described in Livingston and Lewis, 1995), a 4×4 contingency table was created for each content area test and grade level. The $[i,j]$ entry of an accuracy table represents the estimated proportion of students whose true score fell into performance level i and whose observed score fell into performance level j on the Montana CRT. Overall accuracy, which is the proportion of

students whose true and observed performance levels match one another, is the sum of the numbers on the diagonal of the accuracy table.

9.8 Calculating Consistency

To estimate consistency, the true scores are used to estimate the joint distribution of classifications on two independent, parallel test forms. After statistical adjustments (see Livingston and Lewis, 1995), a new 4 × 4 contingency table was created for each test and grade level that shows the proportion of students who would be classified into each performance level by the two (hypothetical) parallel test forms. That is, the [*i*,*j*] entry of a consistency table represents the estimated proportion of students whose observed score on the first form would fall into performance level *i* and whose observed score on the second form would fall into performance level *j*. Overall consistency, which is the proportion of students classified into exactly the same performance level by the two forms of the test, is the sum of the numbers on the diagonal of this new contingency table.

9.9 Kappa

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. Cohen's κ can be used to evaluate the classification consistency of a test from two parallel forms of the test. The two forms in this case were the hypothetical parallel forms used by the Livingston and Lewis method. Because κ is corrected for chance, the values of κ are lower than other consistency estimates.

9.10 Results of Accuracy, Consistency, and Kappa Analyses

Summaries of the Accuracy and Consistency analyses are provided in Tables 9-9 through 9-22. The first section of each table shows the overall accuracy and consistency indices as well as Kappa.

The overall index is, as described above, the sum of the diagonal elements of the appropriate contingency table.

The second section of each table shows accuracy and consistency values conditional upon performance level. In each case, the denominator is the number of students who are associated with a given performance level. For example, the conditional accuracy value is 0.7855 for the *Proficient* category for Grade 4 Math. This indicates that, of the students whose true scores placed them in the *Proficient* category, 78.55% of them would be expected to be in the *Proficient* category if they were categorized according to their observed scores. The corresponding consistency value of .7206 indicates that 72.06% of students with observed scores in the *Proficient* performance level would be expected to score in *Proficient* again if a second, parallel test form were used.

For certain tests, concern may be greatest regarding decisions made about a particular threshold. For example, if a college gave credit to students who achieved an Advanced Placement test score of four or five, but not one, two, or three, one might be interested in the accuracy of the dichotomous decision, below four versus four or above. The third section of the summary tables shows information at each of the cut points. These values indicate the accuracy and consistency of the dichotomous decisions, either above or below the associated cut point. In addition, the false positive and false negative accuracy rates are also provided. These values are estimates of the proportion of students who were categorized above the cut when their true score would place them below the cut (false positive), and vice versa.

Table 9-11: Accuracy and Consistency: Grade 3 MATH

Overall Indices	**Accuracy**		**Consistency**	**Kappa (κ)**
	0.7743		0.6931	0.5589
Indices Conditional on Level			**Accuracy**	**Consistency**
	Novice		0.8229	0.7431
	Nearing Proficiency		0.6121	0.5005
	Proficient		0.7843	0.7251
	Advanced		0.8591	0.7435
Indices at Cut Points		**Accuracy**		**Consistency**
		Accuracy	***False Positives***	***False Negatives***
	N : NP	0.9487	0.0244	0.0269	0.9280
	NP : P	0.9160	0.0449	0.0390	0.8829
	P : A	0.9083	0.0615	0.0302	0.8740

Table 9-12: Accuracy and Consistency: Grade 4 MATH

Overall Indices	**Accuracy**		**Consistency**	**Kappa (κ)**
	0.7842		0.7056	0.5833
Indices Conditional on Level			**Accuracy**	**Consistency**
	Novice		0.8240	0.7467
	Nearing Proficiency		0.6091	0.4975
	Proficient		0.7855	0.7206
	Advanced		0.8818	0.7888
Indices at Cut Points		**Accuracy**		**Consistency**
		Accuracy	***False Positives***	***False Negatives***
	N : NP	0.9474	0.0252	0.0273	0.9263
	NP : P	0.9196	0.0430	0.0374	0.8878
	P : A	0.9159	0.0543	0.0299	0.8830

Table 9-13: Accuracy and Consistency: Grade 5 MATH

Overall Indices	**Accuracy**		**Consistency**	**Kappa (κ)**
	0.7862		0.7068	0.5846
Indices Conditional on Level			**Accuracy**	**Consistency**
	Novice		0.8133	0.7246
	Nearing Proficiency		0.6430	0.5350
	Proficient		0.7782	0.7114
	Advanced		0.8903	0.8018
Indices at Cut Points		**Accuracy**		**Consistency**
		Accuracy	***False Positives***	***False Negatives***
	N : NP	0.9518	0.0223	0.0259	0.9321
	NP : P	0.9172	0.0440	0.0388	0.8845
	P : A	0.9166	0.0540	0.0294	0.8840

Table 9-14: Accuracy and Consistency: Grade 6 MATH

Overall Indices	Accuracy	Consistency	Kappa (κ)
	0.7708	0.6869	0.5661

Indices Conditional on Level		Accuracy	Consistency
	Novice	0.7839	0.6900
	Nearing Proficiency	0.6452	0.5419
	Proficient	0.7572	0.6783
	Advanced	0.8922	0.8054

Indices at Cut Points		Accuracy			Consistency
		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9397	0.0288	0.0315	0.9158
	NP : P	0.9100	0.0497	0.0403	0.8747
	P : A	0.9202	0.0515	0.0283	0.8888

Table 9-15: Accuracy and Consistency: Grade 7 MATH

Overall Indices	Accuracy	Consistency	Kappa (κ)
	0.7643	0.6796	0.5528

Indices Conditional on Level		Accuracy	Consistency
	Novice	0.7571	0.6508
	Nearing Proficiency	0.6228	0.5184
	Proficient	0.7639	0.6860
	Advanced	0.8912	0.8053

Indices at Cut Points		Accuracy			Consistency
		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9352	0.0306	0.0341	0.9095
	NP : P	0.9064	0.0516	0.0420	0.8697
	P : A	0.9215	0.0504	0.0281	0.8906

Table 9-16: Accuracy and Consistency: Grade 8 MATH

Overall Indices	Accuracy	Consistency	Kappa (κ)
	0.7833	0.7026	0.5946

Indices Conditional on Level		Accuracy	Consistency
	Novice	0.7845	0.7050
	Nearing Proficiency	0.6725	0.5712
	Proficient	0.7655	0.6834
	Advanced	0.9063	0.8309

Indices at Cut Points		Accuracy			Consistency
		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9337	0.0342	0.0322	0.9076
	NP : P	0.9178	0.0470	0.0352	0.8853
	P : A	0.9314	0.0439	0.0247	0.9043

Table 9-17: Accuracy and Consistency: Grade 10 MATH

Overall Indices	**Accuracy**	**Consistency**	**Kappa (κ)**	
	0.7840	0.7006	0.5791	
Indices Conditional on Level		**Accuracy**	**Consistency**	
	Novice	0.8044	0.7090	
	Nearing Proficiency	0.7226	0.6345	
	Proficient	0.7805	0.7097	
	Advanced	0.8721	0.7636	
Indices at Cut Points		**Accuracy**		**Consistency**
		Accuracy	***False Positives***	***False Negatives***
	N : NP	0.9490	0.0232	0.0277	0.9283
	NP : P	0.9065	0.0514	0.0421	0.8698
	P : A	0.9284	0.0475	0.0240	0.9001

Table 9-18: Accuracy and Consistency: Grade 3 Reading

Overall Indices	**Accuracy**	**Consistency**	**Kappa (κ)**	
	0.8384	0.7750	0.6399	
Indices Conditional on Level		**Accuracy**	**Consistency**	
	Novice	0.7514	0.6100	
	Nearing Proficiency	0.7236	0.6219	
	Proficient	0.8267	0.7765	
	Advanced	0.8975	0.8340	
Indices at Cut Points		**Accuracy**		**Consistency**
		Accuracy	***False Positives***	***False Negatives***
	N : NP	0.9829	0.0070	0.0102	0.9756
	NP : P	0.9500	0.0246	0.0254	0.9298
	P : A	0.9055	0.0557	0.0388	0.8688

Table 9-19: Accuracy and Consistency: Grade 4 Reading

Overall Indices	**Accuracy**	**Consistency**	**Kappa (κ)**	
	0.8175	0.7461	0.6039	
Indices Conditional on Level		**Accuracy**	**Consistency**	
	Novice	0.7728	0.6500	
	Nearing Proficiency	0.7150	0.6123	
	Proficient	0.8131	0.7585	
	Advanced	0.8746	0.7965	
Indices at Cut Points		**Accuracy**		**Consistency**
		Accuracy	***False Positives***	***False Negatives***
	N : NP	0.9780	0.0094	0.0127	0.9688
	NP : P	0.9420	0.0289	0.0291	0.9187
	P : A	0.8974	0.0612	0.0414	0.8576

Table 9-20: Accuracy and Consistency: Grade 5 Reading

Overall Indices	Accuracy	Consistency	Kappa (κ)
	0.8167	0.7463	0.6064

Indices Conditional on Level		Accuracy	Consistency
	Novice	0.7843	0.6700
	Nearing Proficiency	0.6779	0.5673
	Proficient	0.7817	0.7154
	Advanced	0.8993	0.8369

Indices at Cut Points		Accuracy			Consistency
		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9755	0.0106	0.0139	0.9653
	NP : P	0.9424	0.0287	0.0289	0.9193
	P : A	0.8987	0.0597	0.0416	0.8594

Table 9-21: Accuracy and Consistency: Grade 6 Reading

Overall Indices	Accuracy	Consistency	Kappa (κ)
	0.8127	0.7414	0.5908

Indices Conditional on Level		Accuracy	Consistency
	Novice	0.7942	0.6786
	Nearing Proficiency	0.6448	0.5245
	Proficient	0.7839	0.7204
	Advanced	0.8896	0.8232

Indices at Cut Points		Accuracy			Consistency
		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9771	0.0097	0.0132	0.9675
	NP : P	0.9454	0.0263	0.0283	0.9234
	P : A	0.8899	0.0646	0.0456	0.8473

Table 9-22: Accuracy and Consistency: Grade 7 Reading

Overall Indices	Accuracy	Consistency	Kappa (κ)
	0.8225	0.7544	0.6097

Indices Conditional on Level		Accuracy	Consistency
	Novice	0.7802	0.6571
	Nearing Proficiency	0.6686	0.5532
	Proficient	0.8047	0.7505
	Advanced	0.8939	0.8246

Indices at Cut Points		Accuracy			Consistency
		Accuracy	*False Positives*	*False Negatives*	
	N : NP				
	NP : P	0.9468	0.0260	0.0272	0.9253
	P : A	0.8967	0.0623	0.0410	0.8570

Table 9-23: Accuracy and Consistency: Grade 8 Reading

Overall Indices	Accuracy	Consistency	Kappa (κ)
	0.8111	0.7401	0.6074

Indices Conditional on Level		Accuracy	Consistency
	Novice	0.8250	0.7431
	Nearing Proficiency	0.6366	0.5208
	Proficient	0.7805	0.7119
	Advanced	0.8933	0.8292

Indices at Cut Points		Accuracy			Consistency
		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9690	0.0145	0.0165	0.9563
	NP : P	0.9436	0.0289	0.0276	0.9209
	P : A	0.8981	0.0596	0.0423	0.8588

Table 9-24: Accuracy and Consistency -- Grade 10 Reading

Overall Indices	Accuracy	Consistency	Kappa (κ)
	0.8052	0.7325	0.6005

Indices Conditional on Level		Accuracy	Consistency
	Novice	0.8030	0.6995
	Nearing Proficiency	0.6377	0.5243
	Proficient	0.7930	0.7357
	Advanced	0.8974	0.8204

Indices at Cut Points		Accuracy			Consistency
		Accuracy	*False Positives*	*False Negatives*	
	N : NP	0.9656	0.0151	0.0193	0.9515
	NP : P	0.9309	0.0349	0.0342	0.9035
	P : A	0.9081	0.0581	0.0338	0.8728

CHAPTER 10— SCALING AND EQUATING

The purpose of equating is to ensure that scores obtained from different forms of a test are equivalent to each other. Equating may be used if multiple test forms are administered in the same year, as well as to equate one year's forms to those given in the previous year. Equating ensures that students are not given an unfair advantage or disadvantage because the test form given in one year is easier or harder than the form given in the other year. Once test scores for the forms are placed on an equivalent raw score scale, they then get translated, through the scaling process, to the score scale that is used for reporting. For the 2007 MontCAS, Phase 2 CRT, equating was performed for reading and mathematics, grades 3 through 8 and 10.

10.1 GENERAL RULES

The following general rules are contained in the equating plan for the CRT:

- The goal is to have as many items as possible on the common form constitute the equating set.
- Items used for equating cannot be altered from their appearance in the previous form in any way.
- Whenever possible, items in the equating set should be selected so that they are within five positions of their location on the previous form.
- Passage sets selected for equating should consist of all, or most, of the items associated with the passage.
- The equating set, as a whole group of items, should mirror the characteristics of the common form in terms of content and statistics.

To determine the final set of equating items for each grade level and subject combination, a differential item functioning (DIF) approach using the delta plot method was applied. The 2007 and 2006 *p-values* of each multiple-choice item were transformed to the delta metric. The delta scale is an

inverse normal transformation of percentage correct to a linear scale with a mean of 13 and standard deviation of 4 (Holland & Wainer, 1993). A high delta value indicates a difficult item. For constructed-response items, the average score divided by the maximum possible score, or adjusted p-value, was transformed to the delta metric. The delta values for the potential equating items were computed for each subject in each grade level.

Once all the delta values were calculated for a particular subject and grade, a trend line was fit to the set of points. The perpendicular distance of each item to the regression line was then computed. Items that were not more than three standard deviations away from the regression line were used as equating items. As a result of the delta analyses, a total of ten items was excluded for use as equating items: two from the Grade 8 Mathematics test, and one each from the Grades 3, 4, 6, and 10 Mathematics tests, and the Grades 5, 6, 7 and 8 Reading tests.

10.2 IRT EQUATING

Equating for the MontCAS, Phase 2 CRT used the *anchor-test-nonequivalent-groups* design described by Petersen, Kolen, & Hoover (1989). The fixed common-item IRT procedure was used, in which the anchor items from the previous year's administration were identified during this year's calibrations, and their IRT parameters were fixed to last year's values. This method results in all person and item parameters being on the same θ scale as last year. Because of the equating model that is used for the Montana CRT, the process of equating and scaling does not change the rank ordering of students, give more weight to particular items, or change students' performance-level classifications. Note that the groups of students who took the Montana CRT in 2005-06 and 2006-07 were not equivalent. Item Response Theory (IRT) is particularly useful in equating for nonequivalent groups (Allen & Yen, 1979).

IRT uses mathematical models to define a relationship between an unobserved measure of student ability, usually referred to as theta (θ), and the probability (p) of getting a dichotomous item

correct or of getting a particular score on a polytomous item. In IRT, it is assumed that all items are independent measures of the same construct or ability (i.e., the same θ). There are several IRT models commonly used to specify the relationship between θ and p. For the Montana CRT tests, the 1 parameter logistic (1PL) model was used for multiple-choice and short-answer items and the partial credit model was used for the constructed-response items.

For polytomous items, the generalized partial credit model can be defined as:

$$P_{jk}(\theta) = \frac{\exp \sum_{v=0}^{k} \left[Da_j \left(\theta - b_j + d_v \right) \right]}{\sum_{c=1}^{m} \exp \sum_{v=1}^{c} \left[Da_j \left(\theta - b_j + d_v \right) \right]}$$

where j indexes the items,
k indexes students,
a represents item discrimination,
b represents item difficulty,
d represents category step parameter, and
D is a normalizing constant equal to 1.701.

In the case of the Montana CRT, the a_j term in the above equation is equal to 1.0 for all items. For the dichotomous items, because there is no step parameters (d_v) the above equation reduces to the following:

$$P_j(\theta) = \frac{\exp\left(\theta - b_j\right)}{1 + \exp\left(\theta - b_j\right)}$$

For more information on IRT and IRT models the reader is referred to Hambleton and Swaminathan (1985).

The process of determining the specific mathematical relationship between θ and p is referred to as item calibration. Once items are calibrated, they are defined by a set of parameters which specify a non-linear relationship between θ and p. For more information about item calibration the reader is referred to Lord and Novick (1968) or Hambleton and Swaminathan (1985).

PARSCALE v3.5 (Muraki & Bock, 1999) was the software used to do the IRT analyses. The item parameter files resulting from the analyses are provided in Appendix A. Each item occupied only one block in the calibration run, and the 1.701 normalizing constant was used. A default convergence criterion of 0.001 was used, and all calibrations converged within 32 iterations.

10.3 Translating Raw Scores to Scaled Scores and Performance Levels

Montana CRT scores in each content area are reported on a scale that ranges from 200 to 300. Scaled scores supplement the Montana CRT performance-level results by providing information about the position of a student's results within a performance level. School- and district-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores, or total number of points, on the Montana CRT tests are translated to scaled scores using a data analysis process called *scaling*. Scaling simply converts raw points from one scale to another. In the same way that distance can be expressed in miles or kilometers, or monetary value can be expressed in terms of U.S. dollars or Canadian dollars, student scores on each Montana CRT could be expressed as raw scores (i.e., number right) or scaled scores. It is also important to notice that the raw score to scale score conversion formulae vary from CRT to CRT, analogous to how currency exchange formulae vary from country to country. For example, the scaling conversion formula for Montana's Grade 4 Reading CRT differs from that of the Grade 8 Reading CRT.

It is important to note that converting from raw scores to scaled scores does not change the students' performance-level classifications. Given the relative simplicity of raw scores, it is fair to ask why scaled scores are used in Montana CRT reports instead of raw scores. Foremost, scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Because the standard-setting process typically results in different cut scores across content areas on a raw score basis, it is useful to transform these raw cut scores to a scale that is more easily interpretable and consistent. For the Montana CRT, a score of 225 is the cut score between the

Novice and *Nearing Proficiency* performance levels. This is true regardless of which content area, grade, or year one may be concerned with. If one were to use raw scores, the raw cut score between *Novice* and *Nearing Proficiency* may be, for example, 35 in mathematics at grade 8, but may be 33 in mathematics at grade 10. Using scaled scores greatly simplifies the task of understanding how a student performed.

Cut points for all tests for the MontCAS, Phase 2 CRT were set at standard setting meetings held in June and July, 2006. Cut points were established on the raw score scale, and these raw score cuts were used to determine the scaling coefficients for calculating the scores used for reporting (see description below and Appendix C). Cut points were also determined on the θ-scale. For scaling in 2007, raw score equivalents for these θ-scale cut points were determined using the test characteristic curve (TCC), and these 2007 raw cuts were used to calculate transformation constants.

As previously stated, student scores on the Montana CRT are reported in integer values from 200 to 300 with three scores representing cut scores on each assessment. Two of the three cut points (*Novice*/*Nearing Proficiency* and *Nearing Proficiency*/*Proficient*) were pre-set at 225 and 250, respectively; the third cut point, between *Proficient* and *Advanced*, was allowed to vary across tests, depending on where the raw score cuts were placed. Allowing the upper cut to float results in a single conversion equation for each test; this simplifies interpretation of scaled scores and their summary statistics. Table 10-1 presents the scaled score range for each performance level in each grade/content area combination.

Table 10-1: Scaled Score Range for each Performance Level

Grade	Content Area	*Novice*	*Nearing Proficiency*	*Proficient*	*Advanced*
3	**Reading**	200–224	225–249	250–286	287–300
	Mathematics	200–224	225–249	250–289	290–300
4	**Reading**	200–224	225–249	250–288	289–300
	Mathematics	200–224	225–249	250–290	291–300
5	**Reading**	200–224	225–249	250–286	287–300
	Mathematics	200–224	225–249	250–288	289–300
6	**Reading**	200–224	225–249	250–288	289–300
	Mathematics	200–224	225–249	250–286	287–300
7	**Reading**	200–224	225–249	250–287	288–300
	Mathematics	200–224	225–249	250–288	289–300
8	**Reading**	200–224	225–249	250–288	289–300
	Mathematics	200–224	225–249	250–282	283–300
10	**Reading**	200–224	225–249	250–288	289–300
	Mathematics	200–224	225–249	250–280	281–300

The scaled scores are obtained by a simple linear transformation of the raw scores using the values of 225 and 250 on the scaled score metric and the associated 2007 raw score cut points to define the transformation. The scaling coefficients were calculated using the following formulae:

$$b = 225 - m(x_1)$$

$$b = 250 - m(x_2)$$

$$m = \frac{225 - 250}{x_1 - x_2}$$

where m is the slope of the line providing the relationship between the raw and scaled scores, b is the intercept, x_1 is the cut score on the raw score metric for the *Novice*/*Nearing Proficiency* cut, and x_2 is the cut score on the raw score metric for the *Nearing Proficiency*/*Proficient* cut. Scaled scores were then calculated using the following linear transformation:

$$ss = m(x) + b$$

where x represents a student's raw score. The values obtained using this formula were rounded to the nearest integer and truncated, as necessary, such that no student received a score below 200 or higher than 300. Additional information regarding raw scores, scaled scores, performance level descriptors, and content-specific descriptors may be found in Appendix C.

CHAPTER 11—REPORTING

The CRT assessments were designed to measure student performance against Montana's Content Standards. Consistent with this purpose, results on the CRT were reported in terms of performance levels that describe student performance in relation to these established state standards. There are four performance levels: *Advanced*, *Proficient*, *Nearing Proficiency*, and *Novice* (CRT Performance Level Descriptors, Content-Specific Descriptors, Scaled Score Ranges, and Raw Scores are described in greater detail in Appendix C). Students receive a separate performance-level classification (based on total scaled score) in each content area.

School- and system-level results are reported as the number and percentage of students attaining each performance level at each grade level tested. Disaggregations of students are also reported at the school and system levels. The CRT reports include:

- Student Reports;
- Class Roster & Item-Level Reports;
- School Summary Reports; and
- System Summary Reports.

"Decision Rules" were formulated in early 2007 by OPI and Measured Progress to identify students, during the reporting process, to be excluded from school and system-level reports. A copy of these "Decision Rules" is included in this report as Appendix E.

State summary results were provided to OPI on confidential CDs and via a secure Web site. The report formats are included in Appendix D. Student Reports were delivered to schools on June 29, 2007. All other CRT reporting data were made available to districts and schools online via iAnalyze on June 22, 2007. System Test Coordinators and teachers were also provided with copies of the *Guide to Interpreting the 2007 Criterion-Referenced Test and CRT-ALT Assessment Reports* and *iAnalyze*, to

assist them in understanding the connection between the assessment and the classroom. The guide provides information about the assessment and the use of assessment results.

11.1 iANALYZE

Using advanced Web technology, *iAnalyze* gives Montana educators and administrators the ability to filter data based on test year, grade level, and subject. This allows administrators to isolate test result for specific groups to identify areas of strong or poor performance overall, by content standard or by subgroup. Cross sections of data may be viewed by groupings based on demographics such as gender, Title 1 status, etc.

The confidential nature of the data therein necessitates the strict enforcement of site security. All transmissions are done over Secure Socket Layers (SSL). A system of user role definitions and permissions dictates the scope of access granted to individual users. Organizations (system or school levels) are given administrative power to grant or deny access to their data within the system, and have the ability to specify password durations, disable users, and create custom roles. Personnel using *iAnalyze* may be granted permission to view students' results at an organizational level, or only a select group as defined by the administrator. Each organization is also able to create custom data fields, and import/export functionality is provided. Predefined reports are included in the system, as is the ability to render and print additional copies.

CHAPTER 12—VALIDITY SUMMARY

As stated in the overview chapter, the *Standards for Educational and Psychological Testing* (AERA, et al., 1999) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. The evidence sources around test content, response processes, internal structure, relationship to other variables, and consequences of testing speak to different *aspects* of validity but are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

Evidence on test content validity is meant to determine how well the assessment tasks represent the curriculum and standards for each subject and grade level. Content validation is informed by the item development process, including how the test blueprints and test items align to the curriculum and standards. Viewed through this lens provided by the Standards, evidence based on test content was extensively described in Chapters 2 through 5. Item alignment with Montana content standards; item bias, sensitivity and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training are all components of validity evidence based on test content. As discussed earlier, all CRT test questions are aligned by Montana educators to specific Montana Content Standards, and undergo several rounds of review for content fidelity and appropriateness. Items are presented to students in multiple formats (constructed-response, short-answer and multiple-choice). Finally, tests are administered according to state-mandated standardized procedures, with allowable accommodations, and all test proctors are required to attend annual training sessions.

The scoring information in Chapter 7 describes the steps taken to train and monitor hand-scorers, as well as quality control procedures related to scanning and machine scoring. To speak to

student response processes, however, additional studies would be helpful and might include an investigation of students' cognitive methods using think-aloud protocols.

Evidence based on internal structure is presented in great detail in the discussions of equating and item analyses in Chapters 8 and 9. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlation), differential item functioning analyses, a variety of reliability coefficients, standard errors of measurement, and item response theory parameters and procedures. Each test is equated to the same grade and content test from the prior year in order to preserve the meaning of scores over time. In general, item difficulty and discrimination indices were in acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall.

Evidence based on the consequences of testing is addressed in the scaled scores and reporting information in Chapters 10 and 11, as well as in the test interpretation guide, which is a separate document that is referenced in the discussion of reporting. Each of these chapters speaks to the efforts undertaken to promote accurate and clear information provided to the public regarding test scores. Scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Performance levels provide users with reference points for mastery at each grade level, which is another useful and simple way to interpret scores. Several different standard reports are provided to stakeholders. In addition, a data analysis tool is provided to each school system to allow educators the flexibility to customize reports for local needs. Additional evidence of the consequences of testing could be supplemented with broader investigation of the impact of testing on student learning.

To further support the validation of the assessment program, additional studies might be considered to provide evidence regarding the relationship of CRT results to other variables include the extent to which scores from the CRT assessments converge with other measures of similar constructs, and the extent to which they diverge from measures of different constructs. Relationships among measures of the same or similar constructs can sharpen the meaning of scores and appropriate interpretations by refining the definition of the construct.

The evidence presented in this manual supports inferences of student achievement on the content represented on the Montana Content Standards for Reading and Mathematics for the purposes of program and instructional improvement and as a component of school accountability.

SECTION IV—REFERENCES

Allen, Mary J. & Yen, Wendy M. (1979). *Introduction to Measurement Theory*. Belmont, CA: Wadsworth, Inc.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bock, R. D., and E. Muraki (1999). *PARSCALE: Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software.

Brown, F. G. (1983). *Principles of Educational and Psychological Testing* 3rd ed. Fort Worth, TX: Holt, Rinehart, and Winston.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Dorans, N. J., and P. W. Holland (1993). DIF detection and description. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* pp. 35–66. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., and E. Kulick (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). New York: John Wiley & Sons, Inc

Hambleton, R. K., and W. J. van der Linden (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

Hambleton, R. K., and H. Swaminathan (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer Academic Publishers.

Joint Committee on Testing Practices (2004). *Code of Fair Testing Practices in Education*. Washington, DC: American Psychological Association. Available for download at **http://www.apa.org/science/fairtestcode.html**.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

Lord, F.M., and M. R. Novick (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). *Scaling, Norming, and Equating*. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 221-262).

APPENDIX A: ITEM PARAMETER FILES

Table A-1: Item Parameter Files: Grade 3 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
243048	1	1	-1.4171	0				
247828	1	1	-1.2871	0				
239084	1	1	-0.8979	0				
242811	1	1	-0.631	0				
242985	1	1	0.0335	0				
242804	1	1	-0.976	0				
242712	1	1	-0.7705	0				
242746	1	1	-0.3558	0				
242899	1	1	-0.1992	0				
239013	1	1	-0.6263	0				
247928	1	1	-0.2347	0				
243036	1	1	-1.154	0				
242807	1	1	0.2578	0				
242950	1	1	-0.5801	0				
243031	1	1	-0.2358	0				
242888	1	1	-0.3823	0				
242915	1	1	-1.2756	0				
238998	1	1	0.4767	0				
247926	1	1	-0.7638	0				
242974	1	1	-0.4709	0				
242896	1	1	-0.0746	0				
242722	1	1	-0.8825	0				
247927	1	1	0.7907	0				
239088	1	1	-1.6815	0				
44581	1	1	-1.2623	0				
44564	1	1	-0.7366	0				
44553	1	1	0.305	0				
44569	1	1	-1.3141	0				
44574	1	1	-0.7921	0				
44558	1	1	-1.3934	0				
44546	1	1	-0.5412	0				
44549	1	1	-0.5258	0				
44545	1	1	0.0246	0				
44543	1	1	-0.3196	0				
243010	1	1	-1.592	0				
242742	1	1	-0.8586	0				
50409	1	1	-0.1654	0				
239010	1	1	-0.79	0				
242761	1	1	-0.1343	0				
247902	1	1	-0.4928	0				
242906	1	1	-0.6515	0				

Table A-1: Item Parameter Files: Grade 3 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
242762	1	1	0.3285	0				
243056	1	1	-1.1554	0				
242928	1	1	-0.8348	0				
239009	1	1	-0.1315	0				
34663	1	1	-1.6103	0				
34553	1	1	-1.6779	0				
247975	1	1	-0.2921	0				
242755	1	1	-0.3501	0				
239008	1	1	-0.9038	0				
238999	1	1	0.2772	0				
247961	1	1	-1.0549	0				
247922	1	1	0.1843	0				
243029	1	1	-0.4554	0				
50411	1	1	0.106	0				
242777	1	1	-1.0573	0				
242772	1	1	-0.4574	0				
242776	1	1	-0.4495	0				
63312	4	1	-1.0926	0	0.3708	0.2783	-0.5492	-0.0998
243154	4	1	0.126	0	-0.2533	0.6765	-0.1807	-0.2424

Table A-2: Item Parameter Files: Grade 4 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
242990	1	1	-1.046	0				
243083	1	1	-0.4084	0				
242878	1	1	-0.4508	0				
244323	1	1	-0.1422	0				
243135	1	1	0.0758	0				
248099	1	1	-0.6503	0				
242953	1	1	0.6896	0				
243084	1	1	-1.0206	0				
242873	1	1	-0.0882	0				
248104	1	1	0.4626	0				
243144	1	1	-0.3961	0				
248049	1	1	-0.1477	0				
243172	1	1	0.3974	0				
248073	1	1	0.2307	0				
248067	1	1	-0.5435	0				
248071	1	1	0.3341	0				
244388	1	1	-0.2733	0				
248102	1	1	0.3253	0				
244390	1	1	-0.2402	0				
243037	1	1	-0.4126	0				
242865	1	1	0.111	0				
244352	1	1	-0.6317	0				
243138	1	1	0.0454	0				
248131	1	1	-0.0754	0				
44610	1	1	-0.3199	0				
44607	1	1	-0.4427	0				
44615	1	1	-0.7596	0				
44579	1	1	-0.2531	0				
44617	1	1	0.0908	0				
44608	1	1	-0.6361	0				
44611	1	1	-0.1785	0				
44606	1	1	0.3872	0				
44584	1	1	0.919	0				
44613	1	1	-1.0221	0				
242867	1	1	-0.5746	0				
248100	1	1	-0.2834	0				
248058	1	1	-0.3594	0				
243147	1	1	0.1538	0				
243151	1	1	-0.0521	0				
248007	1	1	0.3974	0				
242908	1	1	-0.4963	0				
244335	1	1	-0.0392	0				
243049	1	1	0.0622	0				

Table A-2: Item Parameter Files: Grade 4 Math

IREF	MAX	A	B	C	D1	D2	D3	D4
248105	1	1	-0.2808	0				
242917	1	1	0.2269	0				
248080	1	1	0.0413	0				
242978	1	1	0.2278	0				
248048	1	1	-0.8135	0				
248132	1	1	0.1909	0				
243082	1	1	0.3002	0				
244361	1	1	-0.5089	0				
243131	1	1	0.0527	0				
248081	1	1	-0.0994	0				
243107	1	1	0.2415	0				
248004	1	1	-0.0794	0				
243173	1	1	-0.4502	0				
243178	1	1	0.2519	0				
246638	1	1	0.3397	0				
246654	4	1	0.5599	0	1.1899	0.4051	-0.2418	-1.3533
246634	4	1	0.2412	0	1.5573	-0.9209	-0.4175	-0.2189

TABLE A-3: ITEM PARAMETER FILES: GRADE 5 MATH

IREF	MAX	A	B	C	D1	D2	D3	D4
237134	1	1	-0.6086	0				
242945	1	1	-0.6269	0				
236166	1	1	0.2486	0				
242926	1	1	-0.8359	0				
242901	1	1	-0.625	0				
243033	1	1	0.2293	0				
242381	1	1	0.2114	0				
237160	1	1	-0.3488	0				
236236	1	1	-0.487	0				
242965	1	1	-0.312	0				
34593	1	1	0.3434	0				
242881	1	1	-0.0278	0				
212921	1	1	0.1299	0				
242882	1	1	0.1385	0				
242367	1	1	-0.6218	0				
242887	1	1	0.3167	0				
237171	1	1	0.2532	0				
242387	1	1	0.0584	0				
236411	1	1	0.2064	0				
242982	1	1	-0.0122	0				
236199	1	1	-0.5286	0				
237155	1	1	-0.5796	0				
242948	1	1	0.4829	0				
242921	1	1	-0.3069	0				
44688	1	1	0.0524	0				
44672	1	1	-0.2242	0				
44695	1	1	-1.2013	0				
44693	1	1	0.303	0				
44696	1	1	-0.7932	0				
44686	1	1	-0.7954	0				
44684	1	1	-1.0835	0				
44690	1	1	0.4877	0				
44681	1	1	-0.295	0				
44674	1	1	-0.7268	0				
242912	1	1	-0.7301	0				
212862	1	1	-0.2546	0				
242902	1	1	-0.6414	0				
243004	1	1	-0.1431	0				
243038	1	1	-0.0054	0				
242363	1	1	-0.7833	0				
236217	1	1	-0.8013	0				
242374	1	1	-0.3888	0				
242986	1	1	-0.4355	0				
242886	1	1	-0.3087	0				

TABLE A-3: ITEM PARAMETER FILES: GRADE 5 MATH

IREF	MAX	A	B	C	D1	D2	D3	D4
242383	1	1	-0.0654	0				
242918	1	1	-0.1966	0				
243026	1	1	-0.866	0				
242910	1	1	-0.6739	0				
243023	1	1	-0.0744	0				
243021	1	1	0.0954	0				
243027	1	1	0.2313	0				
242370	1	1	-0.0373	0				
242372	1	1	-0.0503	0				
242992	1	1	-0.8609	0				
243008	1	1	0.0733	0				
239329	1	1	0.0577	0				
242893	1	1	-1.0092	0				
236017	1	1	0.9541	0				
243015	4	1	-0.3102	0	0.1618	0.009	0.122	-0.2928
242897	4	1	0.3028	0	0.6914	0.4388	-0.6871	-0.4431

TABLE A-4: ITEM PARAMETER FILES: GRADE 6 MATH								
IREF	MAX	A	B	C	D1	D2	D3	D4
243190	1	1	-0.8537	0				
242494	1	1	-0.7317	0				
236747	1	1	-0.4715	0				
243198	1	1	0.2307	0				
243209	1	1	-0.0888	0				
243024	1	1	-0.1675	0				
243128	1	1	-0.1546	0				
243199	1	1	-0.5726	0				
243210	1	1	-0.6302	0				
236869	1	1	-0.2544	0				
243093	1	1	-0.5336	0				
243202	1	1	0.0481	0				
236789	1	1	-0.0129	0				
238479	1	1	-0.3973	0				
243191	1	1	-1.5469	0				
242958	1	1	0.4486	0				
243115	1	1	-0.6974	0				
242966	1	1	0.2933	0				
243193	1	1	0.0851	0				
236876	1	1	0.1721	0				
243047	1	1	0.0762	0				
243136	1	1	0.83	0				
243059	1	1	-0.7923	0				
243122	1	1	-0.0556	0				
44710	1	1	-0.3656	0				
44707	1	1	0.274	0				
44705	1	1	-0.4855	0				
44699	1	1	0.24	0				
44712	1	1	-0.2455	0				
44701	1	1	-0.1378	0				
44702	1	1	0.1676	0				
44703	1	1	-0.1426	0				
44708	1	1	-0.0206	0				
44713	1	1	0.1548	0				
236812	1	1	-0.4476	0				
236659	1	1	-0.323	0				
242499	1	1	-0.0147	0				
243104	1	1	0.3731	0				
239419	1	1	-0.9593	0				
243117	1	1	-0.5217	0				
243203	1	1	-0.3146	0				
243095	1	1	0.0637	0				
239349	1	1	-0.0389	0				
243103	1	1	-1.3909	0				

TABLE A-4: ITEM PARAMETER FILES: GRADE 6 MATH								
IREF	MAX	A	B	C	D1	D2	D3	D4
239345	1	1	-0.306	0				
212907	1	1	0.5162	0				
239772	1	1	-0.3197	0				
243130	1	1	-1.1907	0				
246939	1	1	0.1827	0				
242538	1	1	-1.143	0				
242545	1	1	-1.2339	0				
243100	1	1	0.1628	0				
242983	1	1	0.0661	0				
243112	1	1	0.467	0				
243195	1	1	-0.3471	0				
236715	1	1	-0.2867	0				
239353	1	1	0.2417	0				
239356	1	1	0.1614	0				
242995	4	1	0.4868	0	-0.2864	0.6608	-0.0979	-0.2766
34776	4	1	-0.079	0	-0.4118	1.3002	0.1276	-1.016

TABLE A-5: ITEM PARAMETER FILES: GRADE 7 MATH

IREF	MAX	A	B	C	D1	D2	D3	D4
212991	1	1	-0.8772	0				
245426	1	1	-0.0693	0				
178174	1	1	-0.3131	0				
245114	1	1	-0.1462	0				
244860	1	1	-0.0161	0				
220018	1	1	-0.2885	0				
249907	1	1	-0.7741	0				
178171	1	1	0.3916	0				
244813	1	1	-0.3587	0				
245142	1	1	0.7729	0				
178155	1	1	-0.5556	0				
245108	1	1	0.3069	0				
245477	1	1	0.1579	0				
245117	1	1	0.0907	0				
178167	1	1	-0.5267	0				
245453	1	1	0.1268	0				
249905	1	1	0.2261	0				
245176	1	1	0.3402	0				
178153	1	1	0.1843	0				
245109	1	1	0.2944	0				
245057	1	1	0.9785	0				
245408	1	1	-0.0212	0				
236325	1	1	-0.5643	0				
244864	1	1	-0.7135	0				
44829	1	1	-1.0022	0				
44828	1	1	-0.42	0				
44817	1	1	-0.5943	0				
44795	1	1	0.0002	0				
44802	1	1	0.5561	0				
44804	1	1	-0.4526	0				
44816	1	1	0.1456	0				
44812	1	1	-0.2426	0				
44791	1	1	-0.0273	0				
244863	1	1	-1.4757	0				
244980	1	1	-0.0137	0				
245101	1	1	-0.8174	0				
244868	1	1	-1.001	0				
244826	1	1	0.3558	0				
235987	1	1	0.0696	0				
213088	1	1	-0.102	0				
236509	1	1	-0.3708	0				
178145	1	1	-0.3762	0				
244969	1	1	-0.0724	0				
245100	1	1	-0.189	0				

TABLE A-5: ITEM PARAMETER FILES: GRADE 7 MATH

IREF	MAX	A	B	C	D1	D2	D3	D4
236252	1	1	0.3725	0				
245148	1	1	-0.4171	0				
245140	1	1	-0.0799	0				
245487	1	1	0.0149	0				
245094	1	1	0.4128	0				
178202	1	1	0.1354	0				
245093	1	1	0.2096	0				
245099	1	1	-0.0121	0				
245055	1	1	-0.7429	0				
245200	1	1	-0.2068	0				
236599	1	1	0.1001	0				
236602	1	1	0.9392	0				
236595	1	1	0.6182	0				
249910	4	1	0.0823	0	0.832	0.1967	-0.3104	-0.7182
244982	4	1	0.2358	0	0.0598	0.0622	0.2723	-0.3943

TABLE A-6: ITEM PARAMETER FILES: GRADE 8 MATH

IREF	MAX	A	B	C	D1	D2	D3	D4
212452	1	1	-0.919	0				
248603	1	1	0.0998	0				
243772	1	1	-0.2569	0				
244519	1	1	-0.1896	0				
244461	1	1	-0.2772	0				
244504	1	1	0.1689	0				
244693	1	1	-0.4508	0				
244630	1	1	0.048	0				
244460	1	1	0.3164	0				
244689	1	1	-0.5495	0				
244450	1	1	-0.3573	0				
244489	1	1	0.2102	0				
243334	1	1	-0.01	0				
248631	1	1	0.1404	0				
244488	1	1	0.2887	0				
243497	1	1	-0.1037	0				
52246	1	1	-0.4055	0				
243498	1	1	-0.3495	0				
244563	1	1	0.0043	0				
243793	1	1	0.2542	0				
244577	1	1	0.2129	0				
244473	1	1	-0.096	0				
244566	1	1	0.1197	0				
244568	1	1	0.5382	0				
44648	1	1	-0.1237	0				
44645	1	1	-0.1426	0				
44632	1	1	0.4145	0				
44642	1	1	1.04	0				
44666	1	1	-0.0303	0				
44653	1	1	-0.6327	0				
44662	1	1	0.1362	0				
44631	1	1	-0.1944	0				
44633	1	1	0.1657	0				
44629	1	1	-0.1463	0				
244493	1	1	-1.1166	0				
244518	1	1	-0.7449	0				
244506	1	1	0.3014	0				
244680	1	1	0.0937	0				
212422	1	1	-0.4231	0				
212315	1	1	-0.9101	0				
244552	1	1	-0.4447	0				
244622	1	1	-0.2879	0				
244557	1	1	-0.2426	0				
243782	1	1	0.42	0				

TABLE A-6: ITEM PARAMETER FILES: GRADE 8 MATH

IREF	MAX	A	B	C	D1	D2	D3	D4
244538	1	1	-1.2958	0				
244635	1	1	-0.3917	0				
244453	1	1	0.2187	0				
244528	1	1	0.1766	0				
244596	1	1	0.0338	0				
244567	1	1	0.1475	0				
243343	1	1	-0.2285	0				
212355	1	1	-0.4077	0				
244687	1	1	-0.4084	0				
244595	1	1	0.0811	0				
244564	1	1	0.1859	0				
243774	1	1	0.3368	0				
243712	1	1	0.0579	0				
243770	1	1	-0.5109	0				
244585	4	1	0.2595	0	1.0565	0.3992	-1.3463	-0.1094
244515	4	1	0.223	0	-0.0432	0.7503	-0.6572	-0.0498

TABLE A-7: ITEM PARAMETER FILES: GRADE 10 MATH

IREF	MAX	A	B	C	D1	D2	D3	D4
241018	1	1	-0.9216	0				
243155	1	1	0.2738	0				
241059	1	1	0.464	0				
241082	1	1	0.3372	0				
243076	1	1	-0.8543	0				
243124	1	1	0.3812	0				
241038	1	1	0.6352	0				
243050	1	1	-0.3055	0				
240993	1	1	-0.2968	0				
243096	1	1	-0.2536	0				
249074	1	1	-0.6918	0				
243094	1	1	0.2782	0				
240990	1	1	0.0682	0				
243161	1	1	0.4635	0				
243053	1	1	0.6859	0				
248829	1	1	0.1002	0				
249072	1	1	0.0709	0				
243162	1	1	-0.3321	0				
249103	1	1	0.4215	0				
249038	1	1	-1.1016	0				
248809	1	1	0.1718	0				
243089	1	1	-1.4052	0				
243165	1	1	0.3944	0				
243110	1	1	-0.2868	0				
44572	1	1	-0.3228	0				
44577	1	1	-0.2395	0				
44531	1	1	-0.1074	0				
44583	1	1	-0.1937	0				
44567	1	1	0.1639	0				
44560	1	1	-0.3412	0				
44593	1	1	-1.401	0				
44552	1	1	-0.1772	0				
44573	1	1	-0.0787	0				
44592	1	1	-1.2413	0				
44590	1	1	-0.1127	0				
44587	1	1	-0.4172	0				
44585	1	1	0.3804	0				
44539	1	1	0.5697	0				
44588	1	1	-0.0726	0				
243090	1	1	-0.7267	0				
212572	1	1	0.0413	0				
243114	1	1	-0.2829	0				
240989	1	1	-0.5107	0				
248852	1	1	0.367	0				

TABLE A-7: ITEM PARAMETER FILES: GRADE 10 MATH

IREF	MAX	A	B	C	D1	D2	D3	D4
243141	1	1	0.5929	0				
240999	1	1	-0.1216	0				
243158	1	1	-0.2643	0				
241058	1	1	0.5178	0				
242987	1	1	-0.6818	0				
243149	1	1	0.2411	0				
241048	1	1	0.4719	0				
243140	1	1	0.2952	0				
241089	1	1	0.6709	0				
249045	1	1	-0.1999	0				
243022	1	1	0.9559	0				
241103	1	1	0.0751	0				
243087	1	1	0.0429	0				
51677	1	1	0.2544	0				
242989	1	1	-0.1913	0				
212581	1	1	-1.3317	0				
241196	1	1	-0.0029	0				
241201	1	1	0.9685	0				
241199	1	1	-0.2863	0				
243043	4	1	0.7296	0	-0.6062	0.7004	0.0532	-0.1474
243129	4	1	0.2991	0	0.3515	0.6591	-0.5407	-0.47

TABLE A-8: ITEM PARAMETER FILES: GRADE 3 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
231092	1	1	-0.5774	0				
247848	1	1	-0.8728	0				
231094	1	1	-0.7659	0				
231100	1	1	-0.5368	0				
231099	1	1	-1.6335	0				
243271	1	1	-0.7295	0				
243233	1	1	0.2125	0				
247930	1	1	-1.1026	0				
243243	1	1	-0.8666	0				
247936	1	1	-0.7528	0				
181304	1	1	-0.3501	0				
183910	1	1	0.1608	0				
181297	1	1	-0.2812	0				
181296	1	1	-0.1991	0				
183917	1	1	-0.3783	0				
181299	1	1	-0.3391	0				
183925	1	1	-0.7122	0				
181302	1	1	-0.5644	0				
183918	1	1	-1.2923	0				
181305	1	1	-0.8681	0				
181306	1	1	-0.6336	0				
32727	1	1	-1.1352	0				
33632	1	1	-0.7262	0				
33654	1	1	-1.0975	0				
32729	1	1	-0.3969	0				
33432	1	1	-0.5956	0				
45490	1	1	-0.2791	0				
45484	1	1	0.1455	0				
45487	1	1	-0.1989	0				
45488	1	1	-1.1501	0				
45489	1	1	-0.5104	0				
33616	1	1	0.0018	0				
33644	1	1	-0.4304	0				
33618	1	1	-0.5149	0				
33427	1	1	-0.6186	0				
33646	1	1	0.2072	0				
247962	1	1	-0.492	0				
243263	1	1	0.011	0				
247966	1	1	-0.3438	0				
231289	1	1	0.1704	0				
244265	1	1	-0.0292	0				
33515	1	1	-0.4873	0				
33422	1	1	-0.8236	0				
33412	1	1	-0.6731	0				

TABLE A-8: ITEM PARAMETER FILES: GRADE 3 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
33426	1	1	-0.9946	0				
32821	1	1	-0.4834	0				
33442	1	1	-0.5742	0				
33622	1	1	0.2214	0				
33431	1	1	-0.3929	0				
33443	1	1	-0.4692	0				
33392	1	1	-0.7485	0				
33393	1	1	0.0646	0				
181314	4	1	0.4769	0	1.5951	0.7241	-0.7959	-1.5233
33363	4	1	0.4137	0	2.1247	-0.3504	-0.6681	-1.1062

TABLE A-9: ITEM PARAMETER FILES: GRADE 4 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
235563	1	1	-0.46	0				
248135	1	1	0.1988	0				
244340	1	1	0.0626	0				
248051	1	1	-1.2083	0				
235573	1	1	-0.7905	0				
235550	1	1	0.1602	0				
235552	1	1	-0.6303	0				
235557	1	1	0.5709	0				
244327	1	1	-1.245	0				
235556	1	1	-0.0284	0				
235853	1	1	-0.5577	0				
235857	1	1	-0.0684	0				
244370	1	1	-0.5636	0				
235872	1	1	0.0846	0				
248085	1	1	-0.3132	0				
235879	1	1	0.1674	0				
244329	1	1	-0.7071	0				
244368	1	1	0.2579	0				
235862	1	1	0.0759	0				
235888	1	1	-0.5843	0				
235893	1	1	0.117	0				
45256	1	1	0.0033	0				
45257	1	1	-0.2554	0				
45258	1	1	-0.4473	0				
45259	1	1	-0.3725	0				
45260	1	1	0.1048	0				
45285	1	1	0.291	0				
45286	1	1	-0.3052	0				
45287	1	1	-0.318	0				
45288	1	1	-0.3958	0				
45290	1	1	0.4453	0				
235833	1	1	-0.3722	0				
235836	1	1	-0.1506	0				
235843	1	1	-1.051	0				
235838	1	1	-0.6177	0				
244298	1	1	0.0904	0				
211149	1	1	0.1434	0				
211151	1	1	-0.127	0				
211138	1	1	-0.3976	0				
211144	1	1	-0.1033	0				
211142	1	1	-0.382	0				
235667	1	1	-0.3882	0				
244296	1	1	-0.1015	0				
235671	1	1	0.0615	0				

TABLE A-9: ITEM PARAMETER FILES: GRADE 4 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
244300	1	1	-0.3832	0				
235681	1	1	-0.3015	0				
244377	1	1	-0.969	0				
235712	1	1	-0.4499	0				
235709	1	1	-0.3897	0				
235714	1	1	0.4632	0				
244382	1	1	-0.4128	0				
253329	1	1	0.2993	0				
246688	4	1	0.5428	0	1.2203	0.2257	-0.4511	-0.9949
235720	4	1	0.5016	0	1.5986	0.4886	-0.4615	-1.6258

TABLE A-10: ITEM PARAMETER FILES: GRADE 5 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
213210	1	1	-0.5261	0				
181392	1	1	-1.4457	0				
213209	1	1	-0.3637	0				
181402	1	1	-0.8082	0				
181399	1	1	-0.8733	0				
244738	1	1	-0.6676	0				
212692	1	1	-0.0042	0				
212687	1	1	-0.8231	0				
50094	1	1	-0.5369	0				
212695	1	1	-1.6519	0				
238900	1	1	-0.0212	0				
254079	1	1	-0.5836	0				
238906	1	1	-1.6789	0				
238907	1	1	-0.7554	0				
254083	1	1	-0.0593	0				
238914	1	1	-0.4193	0				
238915	1	1	-0.886	0				
238919	1	1	-1.3702	0				
246651	1	1	-0.4857	0				
238917	1	1	-0.4737	0				
238921	1	1	-0.0402	0				
45054	1	1	-0.7365	0				
45055	1	1	-0.0167	0				
45057	1	1	-1.0903	0				
45058	1	1	-0.3941	0				
50173	1	1	0.2479	0				
45045	1	1	-0.6263	0				
45047	1	1	-0.1511	0				
45048	1	1	-1.1327	0				
45049	1	1	-0.3191	0				
45050	1	1	-0.9964	0				
231124	1	1	-0.8748	0				
231121	1	1	-1.0819	0				
231125	1	1	-1.1924	0				
50096	1	1	-0.4345	0				
231129	1	1	-0.614	0				
231219	1	1	0.2031	0				
231227	1	1	-0.2972	0				
231228	1	1	-0.9172	0				
231233	1	1	-1.0158	0				
231234	1	1	-0.3699	0				
231144	1	1	-0.4968	0				
231149	1	1	-0.853	0				
231148	1	1	-0.4216	0				

TABLE A-10: ITEM PARAMETER FILES: GRADE 5 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
231150	1	1	-0.3046	0				
244727	1	1	-0.8675	0				
231152	1	1	-0.3734	0				
231155	1	1	0.1067	0				
231158	1	1	0.4162	0				
231160	1	1	-1.545	0				
231167	1	1	-1.0171	0				
231171	1	1	-0.6583	0				
238931	4	1	0.2647	0	1.7677	0.5549	-0.7991	-1.5235
231173	4	1	0.2183	0	0.3872	0.608	-0.2021	-0.7931

TABLE A-11: ITEM PARAMETER FILES: GRADE 6 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
231449	1	1	-1.8105	0				
254074	1	1	-0.7223	0				
231444	1	1	-0.0432	0				
231451	1	1	-0.9054	0				
231447	1	1	-0.7215	0				
238762	1	1	-0.4233	0				
254030	1	1	-0.9507	0				
238898	1	1	-0.9418	0				
238895	1	1	-0.7885	0				
238768	1	1	-0.283	0				
253993	1	1	-0.0016	0				
254004	1	1	-0.1812	0				
254075	1	1	-0.4252	0				
246565	1	1	-0.3265	0				
246936	1	1	-0.9629	0				
231380	1	1	-0.827	0				
246590	1	1	-0.3373	0				
246938	1	1	0.2056	0				
254007	1	1	-0.5216	0				
254008	1	1	0.118	0				
231393	1	1	-0.7587	0				
33570	1	1	0.1697	0				
33831	1	1	-0.9325	0				
33813	1	1	-0.8979	0				
32902	1	1	-0.713	0				
32903	1	1	-0.3631	0				
171149	1	1	-0.1193	0				
50720	1	1	-0.6382	0				
171153	1	1	-0.7322	0				
171154	1	1	-0.1486	0				
171155	1	1	-0.9945	0				
231410	1	1	-1.2417	0				
246585	1	1	-1.3616	0				
231403	1	1	0.0629	0				
246587	1	1	0.4106	0				
231409	1	1	-0.9757	0				
231494	1	1	-0.4599	0				
231489	1	1	-0.3729	0				
254059	1	1	-0.233	0				
231492	1	1	-0.1893	0				
231498	1	1	-0.8082	0				
254039	1	1	-1.0961	0				
246581	1	1	-1.3692	0				
231424	1	1	-0.8467	0				

TABLE A-11: ITEM PARAMETER FILES: GRADE 6 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
246601	1	1	-0.466	0				
231428	1	1	-0.6507	0				
246588	1	1	-1.0721	0				
254049	1	1	-0.2892	0				
231425	1	1	-0.7469	0				
231423	1	1	-0.489	0				
231430	1	1	0.0306	0				
231429	1	1	-0.7024	0				
242230	4	1	-0.1308	0	1.4905	0.738	-1.1479	-1.0807
231432	4	1	0.2526	0	1.5217	0.0107	-0.4684	-1.064

TABLE A-12: ITEM PARAMETER FILES: GRADE 7 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
245551	1	1	0.0098	0				
212130	1	1	-0.3336	0				
212133	1	1	-0.501	0				
212135	1	1	-0.156	0				
212136	1	1	-0.9374	0				
249692	1	1	-0.5943	0				
212118	1	1	-0.516	0				
212120	1	1	-0.3872	0				
254403	1	1	-0.3111	0				
254405	1	1	-0.5535	0				
26049	1	1	-0.7066	0				
50368	1	1	-0.3317	0				
50395	1	1	-0.5658	0				
50370	1	1	-0.7672	0				
50371	1	1	0.0093	0				
50372	1	1	-0.8532	0				
50376	1	1	-0.6983	0				
50377	1	1	-0.4665	0				
50378	1	1	-0.5829	0				
26052	1	1	-0.7135	0				
50379	1	1	-0.859	0				
171353	1	1	-1.2574	0				
171354	1	1	-1.0456	0				
171351	1	1	-0.6155	0				
171358	1	1	-0.6493	0				
171360	1	1	-0.8063	0				
171331	1	1	-0.3862	0				
171333	1	1	-0.1647	0				
171334	1	1	-0.7986	0				
171335	1	1	0.1793	0				
171336	1	1	-0.6207	0				
244959	1	1	-0.7887	0				
212210	1	1	-0.2497	0				
249691	1	1	-0.854	0				
212212	1	1	-0.5527	0				
249699	1	1	-0.6157	0				
254408	1	1	-0.465	0				
212171	1	1	-0.1668	0				
212170	1	1	-0.6051	0				
212166	1	1	-0.726	0				
212174	1	1	-0.5976	0				
249848	1	1	-0.5968	0				
245435	1	1	-0.0875	0				
249852	1	1	-0.6283	0				

TABLE A-12: ITEM PARAMETER FILES: GRADE 7 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
244962	1	1	0.2933	0				
238711	1	1	0.3774	0				
239781	1	1	-0.5341	0				
245092	1	1	-1.2132	0				
238710	1	1	-0.3302	0				
238722	1	1	-0.3927	0				
249886	1	1	-0.2114	0				
249854	1	1	-0.6266	0				
26054	4	1	-0.2554	0	1.415	0.5775	-0.5843	-1.4081
33781	4	1	-0.1348	0	1.4492	0.399	-0.5821	-1.2661

TABLE A-13: ITEM PARAMETER FILES: GRADE 8 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
235944	1	1	0.0379	0				
235945	1	1	0.0422	0				
248592	1	1	-0.7365	0				
235946	1	1	0.0065	0				
235953	1	1	-0.2333	0				
236261	1	1	-0.5616	0				
248626	1	1	-0.2774	0				
236276	1	1	-0.6021	0				
236280	1	1	0.0312	0				
248627	1	1	-1.2593	0				
236084	1	1	-0.1746	0				
248772	1	1	-1.1112	0				
236092	1	1	-0.647	0				
248778	1	1	-0.1767	0				
248780	1	1	-0.7551	0				
236103	1	1	-0.919	0				
236105	1	1	-0.876	0				
248782	1	1	-0.3856	0				
244497	1	1	-0.5291	0				
236108	1	1	0.0887	0				
236127	1	1	-0.4972	0				
44884	1	1	-1.4502	0				
44886	1	1	-0.9057	0				
44888	1	1	-0.5315	0				
44891	1	1	-0.3017	0				
44895	1	1	-1.5541	0				
33184	1	1	-0.2667	0				
33183	1	1	-0.8546	0				
33185	1	1	-0.8217	0				
33188	1	1	-0.4387	0				
33190	1	1	-1.0698	0				
236001	1	1	-0.1795	0				
236005	1	1	-0.1754	0				
236025	1	1	-0.2928	0				
236032	1	1	-0.6902	0				
236035	1	1	-0.2928	0				
236328	1	1	-0.1646	0				
236331	1	1	-0.1364	0				
248646	1	1	-0.3341	0				
236348	1	1	-0.2345	0				
236350	1	1	-0.7336	0				
248796	1	1	-0.5121	0				
248799	1	1	-0.484	0				
244589	1	1	-0.3646	0				

TABLE A-13: ITEM PARAMETER FILES: GRADE 8 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
236420	1	1	-0.1959	0				
248805	1	1	-0.1512	0				
236438	1	1	0.034	0				
236435	1	1	-0.4143	0				
236419	1	1	-0.861	0				
236430	1	1	-0.0955	0				
248794	1	1	0.1309	0				
248811	1	1	-0.4211	0				
248785	4	1	0.5788	0	1.1107	0.598	-0.6199	-1.0889
248816	4	1	0.5709	0	1.3678	0.4037	-0.625	-1.1465

TABLE A-14: ITEM PARAMETER FILES: GRADE 10 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
249035	1	1	-0.275	0				
235679	1	1	-0.138	0				
235680	1	1	-0.3769	0				
235682	1	1	-0.9119	0				
235684	1	1	-0.5228	0				
248826	1	1	-0.3552	0				
248828	1	1	-0.7782	0				
235860	1	1	-0.3252	0				
248833	1	1	-0.412	0				
235864	1	1	-0.0567	0				
235690	1	1	0.3411	0				
235692	1	1	-0.7824	0				
235695	1	1	-0.412	0				
235696	1	1	-0.795	0				
235699	1	1	-0.0131	0				
235706	1	1	-0.5342	0				
235708	1	1	-0.535	0				
235703	1	1	-0.5556	0				
235710	1	1	-0.3151	0				
235713	1	1	-0.6265	0				
248706	1	1	-0.097	0				
44551	1	1	-0.158	0				
44557	1	1	0.3503	0				
44561	1	1	-0.2608	0				
44568	1	1	-0.3632	0				
44566	1	1	0.0801	0				
44555	1	1	-0.7524	0				
44562	1	1	0.5467	0				
44570	1	1	0.5112	0				
44534	1	1	-1.3767	0				
44536	1	1	-1.6147	0				
44537	1	1	-0.4729	0				
44538	1	1	-0.4239	0				
44540	1	1	-0.4272	0				
44541	1	1	-0.0941	0				
44548	1	1	-1.0725	0				
235874	1	1	-1.189	0				
246526	1	1	-0.7272	0				
249040	1	1	-0.1669	0				
235885	1	1	-0.2443	0				
249041	1	1	-0.4931	0				
235657	1	1	-1.1603	0				
235659	1	1	-0.2705	0				
235661	1	1	0.4039	0				

TABLE A-14: ITEM PARAMETER FILES: GRADE 10 READING

IREF	MAX	A	B	C	D1	D2	D3	D4
249046	1	1	0.0078	0				
235666	1	1	0.2107	0				
235723	1	1	-0.5813	0				
235726	1	1	-1.1354	0				
235729	1	1	-0.8275	0				
235730	1	1	-0.5307	0				
248710	1	1	-0.1435	0				
235735	1	1	0.3481	0				
248713	1	1	-0.0841	0				
235741	1	1	-0.8955	0				
248714	1	1	0.2692	0				
235744	1	1	-0.3592	0				
235734	1	1	-0.3779	0				
235719	4	1	0.3631	0	1.4729	0.2155	-0.5889	-1.0995
248718	4	1	0.2476	0	1.5157	0.522	-0.638	-1.3996

APPENDIX B: TECHNICAL ADVISORY COMMITTEE

TABLE B-1: 2007 TECHNICAL ADVISORY COMMITTEE (TAC) MEMBERS

First Name	Last Name	Position	Department	Organization
Art	Bangert, Ph.D.	Assistant Professor	Adult and Higher Education	Montana State University
Susan	Brookhart, Ph.D.	President		Brookhart Enterprises, LLC
Ellen	Forte, Ph.D.	President		edCount, LLC
Michael	Kozlow, Ph.D.	Program Director	Assessment Program	
Scott	Marion, Ph.D.	Vice-President		Center for Assessment
Stanley	Rabinowitz, Ph.D.	Program Director	Assessment & Standards Development Services	WestEd
Derek	Briggs, Ph.D.	Assistant Professor	School of Education	University of Colorado

APPENDIX C: CRT PERFORMANCE LEVEL DESCRIPTORS, SCALED SCORES, AND RAW SCORES

TABLE C-1: CRT PERFORMANCE LEVEL DESCRIPTORS (GENERAL)

Advanced	This level denotes superior performance.
Proficient	This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
Nearing Proficiency	This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.
Novice	This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

TABLE C-2: CRT SCALED SCORE RANGES FOR PERFORMANCE LEVELS: GRADE 3

	Reading	Mathematics
Advanced	287-300	290-300
Proficient	250-286	250-289
Nearing Proficiency	225-249	225-249
Novice	200-224	200-224

TABLE C-3: CRT SCALED SCORE RANGES FOR PERFORMANCE LEVELS: GRADE 4

	Reading	Mathematics
Advanced	289-300	291-300
Proficient	250-288	250-290
Nearing Proficiency	225-249	225-249
Novice	200-224	200-224

TABLE C-4: CRT SCALED SCORE RANGES FOR PERFORMANCE LEVELS: GRADE 5

	Reading	Mathematics
Advanced	287-300	289-300
Proficient	250-286	250-288
Nearing Proficiency	225-249	225-249
Novice	200-224	200-224

TABLE C-5: CRT SCALED SCORE RANGES FOR PERFORMANCE LEVELS: GRADE 6

	Reading	Mathematics
Advanced	289-300	287-300
Proficient	250-288	250-286
Nearing Proficiency	225-249	225-249
Novice	200-224	200-224

TABLE C-6: CRT SCALED SCORE RANGES FOR PERFORMANCE LEVELS: GRADE 7

	Reading	Mathematics
Advanced	288-300	289-300
Proficient	250-287	250-288
Nearing Proficiency	225-249	225-249
Novice	200-224	200-224

TABLE C-7: CRT SCALED SCORE RANGES FOR PERFORMANCE LEVELS: GRADE 8

	Reading	Mathematics
Advanced	289-300	283-300
Proficient	250-288	250-282
Nearing Proficiency	225-249	225-249
Novice	200-224	200-224

TABLE C-8: CRT SCALED SCORE RANGES FOR PERFORMANCE LEVELS: GRADE 10

	Reading	Mathematics
Advanced	289-300	281-300
Proficient	250-288	250-280
Nearing Proficiency	225-249	225-249
Novice	200-224	200-224

TABLE C-9: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 3 READING

Proficiency Level	Raw Score Range	% in Level
Advanced	44-60	37.8
Proficient	28-43	46.2
Nearing Proficiency	18-27	12.6
Novice	0-17	3.5

TABLE C-10: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 4 READING

Proficiency Level	Raw Score Range	% in Level
Advanced	45-60	32.8
Proficient	30-44	47.4
Nearing Proficiency	20-29	14.8
Novice	0-19	5.0

TABLE C-11: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 5 READING

Proficiency Level	Raw Score Range	% in Level
Advanced	45-60	41.4
Proficient	32-44	39.9
Nearing Proficiency	23-31	12.7
Novice	0-22	5.9

TABLE C-12: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 6 READING

Proficiency Level	Raw Score Range	% in Level
Advanced	45-60	41.1
Proficient	32-44	42.2
Nearing Proficiency	24-31	10.9
Novice	0-23	5.8

TABLE C-13: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 7 READING

Proficiency Level	Raw Score Range	% in Level
Advanced	46-60	38.3
Proficient	31-45	45.5
Nearing Proficiency	22-30	11.1
Novice	0-21	5.1

TABLE C-14: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 8 READING

Proficiency Level	Raw Score Range	% in Level
Advanced	47-60	38.2
Proficient	34-46	40.9
Nearing Proficiency	26-33	11.2
Novice	0-25	9.6

TABLE C-15: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 10 READING

Proficiency Level	Raw Score Range	% in Level
Advanced	51-65	35.4
Proficient	38-50	43.1
Nearing Proficiency	30-37	12.0
Novice	0-29	9.4

TABLE C-16: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 3 MATH

Proficiency Level	Raw Score Range	% in Level
Advanced	55-66	22.6
Proficient	42-54	45.3
Nearing Proficiency	34-41	17.4
Novice	0-33	14.7

TABLE C-17: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 4 MATH

Proficiency Level	Raw Score Range	% in Level
Advanced	50-66	26.0
Proficient	36-49	41.9
Nearing Proficiency	28-35	16.9
Novice	0-27	15.3

TABLE C-18: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 5 MATH

Proficiency Level	Raw Score Range	% in Level
Advanced	48-66	28.0
Proficient	34-47	39.1
Nearing Proficiency	25-33	19.4
Novice	0-24	13.5

TABLE C-19: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 6 MATH

Proficiency Level	Raw Score Range	% in Level
Advanced	46-66	25.9
Proficient	33-45	37.8
Nearing Proficiency	24-32	22.2
Novice	0-23	14.0

TABLE C-20: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 7 MATH

Proficiency Level	Raw Score Range	% in Level
Advanced	43-65	26.6
Proficient	30-42	38.1
Nearing Proficiency	22-29	21.3
Novice	0-21	14.0

TABLE C-21: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 8 MATH

Proficiency Level	Raw Score Range	% in Level
Advanced	46-66	26.9
Proficient	32-45	33.0
Nearing Proficiency	22-31	23.3
Novice	0-21	16.8

TABLE C-22: RAW SCORE RANGE AND PERCENT OF STUDENTS IN EACH PERFORMANCE LEVEL: GRADE 10 MATH

Proficiency Level	Raw Score Range	% in Level
Advanced	48-71	21.2
Proficient	33-47	33.9
Nearing Proficiency	21-32	32.2
Novice	0-20	12.7

APPENDIX D: REPORT SHELLS

Student Report

Class Roster & Item-Level Report

School Summary Report

System Summary Report

CRT Performance Level Descriptors

The Performance Level Descriptors below describe students' knowledge, skills, and abilities in a content area. These descriptions provide a picture or profile of student achievement at the four performance levels: **Advanced**, **Proficient**, **Nearing Proficiency**, and **Novice**. Grade and content performance level descriptors may be found on OPI's web site at http://www.opi.mt.gov/assessment/index.html

Advanced
This level denotes superior performance.

Proficient
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

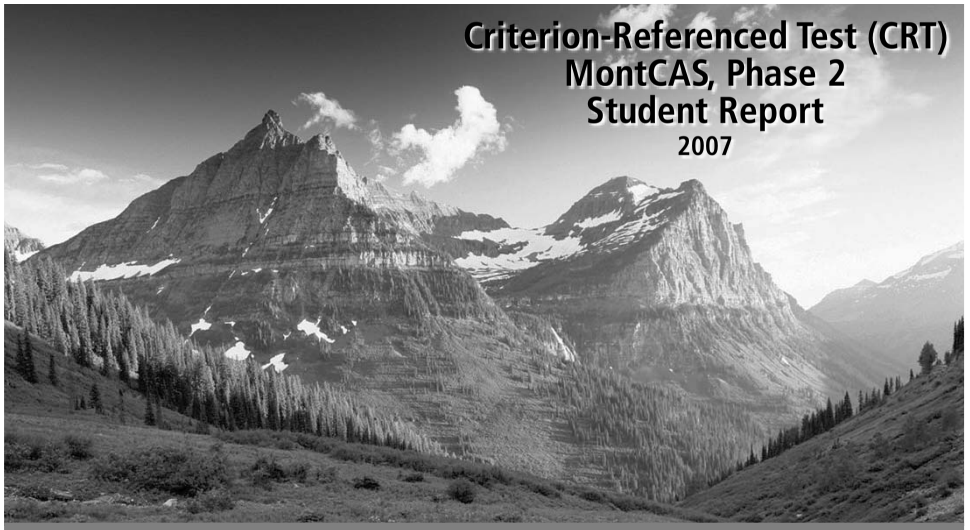
Novice
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

	Score Ranges	
	Reading	**Math**
Advanced	(287-300)	(290-300)
Proficient	(250-286)	(250-289)
Nearing Proficiency	(225-249)	(225-249)
Novice	(200-224)	(200-224)

For more information regarding student assessments in Montana, check out the Office of Public Instruction's Parents Page at http://www.opi.mt.gov/parents.

OPI Contact
Judy Snow, State Assessment Director
406-444-3656
jsnow@mt.gov

Linda McCulloch, Superintendent
Montana Office of Public Instruction
PO Box 202501
Helena, Montana 59620-2501
http://www.opi.mt.gov



Criterion-Referenced Test (CRT)
MontCAS, Phase 2
Student Report
2007

Student Name:
School:
System:
Grade: 03

Dear Parents/Guardians:

This report contains the results of the Spring 2007 Montana Comprehensive Assessment System (MontCAS) Criterion-Referenced Test (CRT) that your child took in March. The CRT provides schools with information to evaluate and improve curriculum and instruction to help all students meet Montana's content standards. This report provides important information about your child's performance on the assessment along with state results.

The CRT contains multiple-choice and short-answer questions. The test measures a student's knowledge of subject matter identified in the Montana State Standards for Reading and Mathematics.

It is important to remember that the CRT is just one measure of your child's academic progress. Your local school staff can provide further information about your child's performance in school. The CRT, which is required by the No Child Left Behind Act, is part of an ongoing statewide educational improvement process. Working together, we can ensure that Montana's children continue to receive a high-quality education.

Sincerely,

Linda McCulloch

Montana Superintendent of Public Instruction

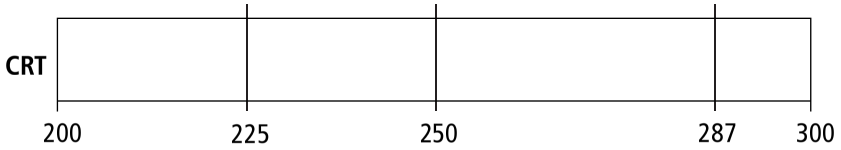
Scaled Scores on the CRT

The criterion-referenced test (CRT) is designed to measure student performance against the learning goals described in the Montana Content Standards (http://www.opi.state.mt.us/standards/index.html). Consistent with this purpose, results on the CRT are reported according to performance levels that describe student performance in relation to the established state standards. There are four performance levels: **Advanced**, **Proficient**, **Nearing Proficiency**, and **Novice**. Your child's performance levels in reading and mathematics are based on a total scaled score in each content area. Scaled scores in each content area range from 200 to 300. Your child's performance levels, based on the scaled scores, are shown in the bar graphs below.

Scaled Scores

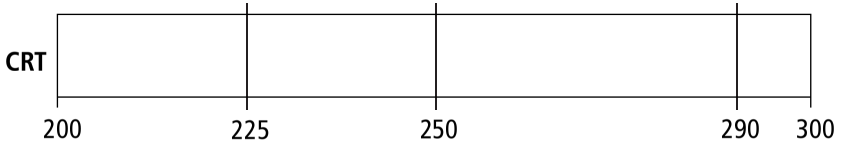
STUDENT RESULTS FOR READING

Performance Level:
Student Scaled Score:



STUDENT RESULTS FOR MATHEMATICS

Performance Level:
Student Scaled Score:



Scores on Montana Content Standards

In addition to performance levels, CRT results are reported for Montana Content Standards in reading and mathematics. Unlike scaled scores which provide a total performance level score, Montana Content Standard Scores provide more specific information about your child's achievement on the CRT. The chart on the following page shows your child's performance in each area of study within subject areas (Montana Content Standards for reading and math). These results can be used to show your child's relative strengths or weaknesses.

Contact your student's school for more information about the following symbols:
† Student did not complete the assessment.
§ Student took non-standard accommodation.

Scores on Montana Standards — Percentage of Points Earned

Reading Standards	Points Possible	Student Percentage	State Percentage
1. Students construct meaning as they comprehend, interpret, and respond to what they read.	23		
2. Students apply a range of skills and strategies to read.	19		
3. Students set goals, monitor, and evaluate their reading progress.	This standard is not measurable in a statewide assessment.		
4. Students select, read, and respond to print and nonprint material for a variety of purposes.	10		
5. Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences.	8		
Math Standards			
1. Problem Solving	8		
2. Numbers and Operations	13		
3. Algebra	6		
4. Geometry	11		
5. Measurement	9		
6. Data Analysis, Statistics, and Probability	12		
7. Patterns, Relations, and Functions	7		

0 25 50 75 100

Student percentage of points earned

State percentage of points earned

MontCAS, Phase 2
CRT

Reading
Roster & Item-Level Report
Confidential

Spring 2007

Class:
School:
System:

Grade: 03
Page: of

Name	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	46	Scaled Score	Perf. Level
Standard	2	2	4	4	4	4	2	2	1	1	1	2	1	4	2	1	2	1	2	4	5	5	1	2	1	1	1	5	1	1	5	1	2		
Correct Response	B	A	C	B	B	D	D	A	C	B	B	B	C	A	D	C	A	B	B	D	C		A	D	B	C	C	C	A	B	D	C	C		
Total Possible Points	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1		
Class Average																																			
School Average																																			
System Average																																			
State Average																																			

† Student did not complete the assessment. § Student took non-standard accommodation. ¥ Not in school and/or system for full academic year. IR = Irregular Test Administration

MontCAS, Phase 2 CRT

Reading
Roster & Item-Level Report
Confidential

Spring 2007

Class:
School:
System:

Grade: 03
Page: of

Name	Item Number	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	Scaled Score	Perf. Level
	Standard	1	1	1	4	2	4	4	2	4	1	2	1	1	1	2	2	1	1	1	5	2		
	Correct Response	B	B	B	C	C	B	A	D	D	B	C	C	A	B	C	D	B	A	C	D			
	Total Possible Points	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4		
Class Average																								
School Average																								
System Average																								
State Average																								

† Student did not complete the assessment. § Student took non-standard accommodation. ¥ Not in school and/or system for full academic year. IR = Irregular Test Administration

MontCAS, Phase 2 CRT

Mathematics
Roster & Item-Level Report
Confidential

Spring 2007

Class:
School:
System:

Grade: 03
Page: of

Name	Item Number	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	35	36	37	Scaled Score	Perf. Level
	Standard	6	5	4	3	5	7	2	2	7	4	5	6	7	4	6	3	4	2	7	5	7	2	6	4	6	7	5	4	5	6	4	1	3		
	Correct Response	B	C	A	A	C	C	B	B	A	C	B	B	D	B	D	D	A	C	C	A	C	D	B	B		A	D	C	C	B	D	B	B		
	Total Possible Points	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1		
Class Average																																				
School Average																																				
System Average																																				
State Average																																				

† Student did not complete the assessment. § Student took non-standard accommodation. ¥ Not in school and/or system for full academic year. IR = Irregular Test Administration

MontCAS, Phase 2 CRT

Mathematics
Roster & Item-Level Report
Confidential

Spring 2007

Class:
School:
System:

Grade: 03
Page: of

Item Number	38	39	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68							Scaled Score	Perf. Level
Standard	1	2	5	2	2	3	1	4	7	1	6	4	3	4	2	5	2	3	2	6	5	6	4	2	2	2	1								
Correct Response	A	C	C	C	C	B	A	B	D	D	A	B	D	A	B	B	C	C	D	C	C	D	B												
Name / Total Possible Points	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4								
Class Average																																			
School Average																																			
System Average																																			
State Average																																			

† Student did not complete the assessment. § Student took non-standard accommodation. ¥ Not in school and/or system for full academic year. IR = Irregular Test Administration

Legend for Roster and Item-Level Report

Mathematics and Reading

Item Number: This is the number of the question on the test.

Standard: This shows the standard each question correlates with.

Total Possible Points: This number indicates the total possible points awarded for the item.

Name: Each student's name is listed, followed by response information for each item on the test.

For multiple choice questions, a plus sign (+) indicates that the student selected the correct response. If the student answered incorrectly, the letter of their response is indicated. A space indicates that the student made no selection.

For all open-response items, a number (0, 1, 2, 3, or 4) indicates how many points the student earned for that item. B indicates that the student left the question blank.

Summary of Scores: Averages are listed for various groups of students (e.g. school and system).

For all items, the average of the number of points awarded to all students in that group is shown.

Scaled Score: This column shows the scaled score that corresponds to the total points earned on items that correlate to Montana Standards.

Performance Level: This column shows the performance level into which the student's scores fall.

Advanced (A) This level denotes superior performance.

Proficient (P) This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (NP) This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (N) This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2
CRT

System:
Grade: 03
Spring 2007

Reading

System Summary Report

I. Distribution of Scores

Perf. Level	Scores	System			State		
		Number	% of Students	% of Students in Cat.	Number	% of Students	% of Students in Cat.
Advanced	298-300						
	295-297						
	293-294						
	290-292						
	287-289						
Proficient	280-286						
	272-279						
	265-271						
	257-264						
	250-256						
Nearing Proficiency	245-249						
	240-244						
	235-239						
	230-234						
	225-229						
Novice	220-224						
	215-219						
	210-214						
	205-209						
	200-204						

II. Subtest Results

Reading		Possible Points	Average Points Earned	
			System	State
Total Points		60		
Standards	1. Students construct meaning as they comprehend, interpret, and respond to what they read	23		
	2. Students apply a range of skills and strategies to read	19		
	3. Students set goals, monitor, and evaluate their reading progress	This standard is not measurable in a statewide assessment.		
	4. Students select, read, and respond to print and nonprint material for a variety of purposes	10		
	5. Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	8		

CRT Performance Level Descriptors

Advanced (287-300)
This level denotes superior performance.

Proficient (250-286)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2
CRT

System:
Grade: 03
Spring 2007

Mathematics

System Summary Report

I. Distribution of Scores

Perf. Level	Scores	System			State		
		Number	% of Students	% of Students in Cat.	Number	% of Students	% of Students in Cat.
Advanced	299-300						
	297-298						
	294-296						
	292-293						
	290-291						
Proficient	282-289						
	274-281						
	266-273						
	258-265						
	250-257						
Nearing Proficiency	245-249						
	240-244						
	235-239						
	230-234						
	225-229						
Novice	220-224						
	215-219						
	210-214						
	205-209						
	200-204						

II. Subtest Results

Mathematics		Possible Points	Average Points Earned	
			System	State
Total Points		66		
Standards	1. Problem Solving	8		
	2. Numbers and Operations	13		
	3. Algebra	6		
	4. Geometry	11		
	5. Measurement	9		
	6. Data Analysis, Statistics, and Probability	12		
	7. Patterns, Relations, and Functions	7		

CRT Performance Level Descriptors

Advanced (290-300)
This level denotes superior performance.

Proficient (250-289)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2
CRT

School:
System:
Grade: 03
Spring 2007

Reading

School Summary Report

I. Distribution of Scores

Perf. Level	Scores	School			System			State		
		N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.
Advanced	298-300									
	295-297									
	293-294									
	290-292									
	287-289									
Proficient	280-286									
	272-279									
	265-271									
	257-264									
	250-256									
Nearing Proficiency	245-249									
	240-244									
	235-239									
	230-234									
	225-229									
Novice	220-224									
	215-219									
	210-214									
	205-209									
	200-204									

II. Subtest Results

Reading		Possible Points	Average Points Earned		
			School	System	State
Total Points		60			
Standards	1. Students construct meaning as they comprehend, interpret, and respond to what they read	23			
	2. Students apply a range of skills and strategies to read	19			
	3. Students set goals, monitor, and evaluate their reading progress	This standard is not measurable in a statewide assessment.			
	4. Students select, read, and respond to print and nonprint material for a variety of purposes	10			
	5. Students gather, analyze, synthesize, and evaluate information from a variety of sources, and communicate their findings in ways appropriate for their purposes and audiences	8			

CRT Performance Level Descriptors

Advanced (287-300)
This level denotes superior performance.

Proficient (250-286)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2
CRT

School:
System:
Grade: 03
Spring 2007

Mathematics

School Summary Report

I. Distribution of Scores

Perf. Level	Scores	School			System			State		
		N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.	N	% of Students	% of Students in Cat.
Advanced	299-300									
	297-298									
	294-296									
	292-293									
	290-291									
Proficient	282-289									
	274-281									
	266-273									
	258-265									
	250-257									
Nearing Proficiency	245-249									
	240-244									
	235-239									
	230-234									
	225-229									
Novice	220-224									
	215-219									
	210-214									
	205-209									
	200-204									

II. Subtest Results

Mathematics		Possible Points	Average Points Earned		
			School	System	State
Total Points		66			
Standards	1. Problem Solving	8			
	2. Numbers and Operations	13			
	3. Algebra	6			
	4. Geometry	11			
	5. Measurement	9			
	6. Data Analysis, Statistics, and Probability	12			
	7. Patterns, Relations, and Functions	7			

CRT Performance Level Descriptors

Advanced (290-300)
This level denotes superior performance.

Proficient (250-289)
This level denotes solid academic performance for each benchmark. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Nearing Proficiency (225-249)
This level denotes that the student has partial mastery or prerequisite knowledge and skills fundamental for proficient work at each benchmark.

Novice (200-224)
This level denotes that the student is beginning to attain the prerequisite knowledge and skills that are fundamental for work at each benchmark.

MontCAS, Phase 2 CRT

Confidential

Reading **System Summary Report**

System:
Grade: 03
Spring 2007

III. Results for Subgroups of Students

Reporting Category	System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students										
Gender										
Male										
Female										
Ethnicity										
American Indian or Alaska Native										
Asian										
Hispanic										
Black or African American										
Native Hawaiian or Other Pacific Islander										
White										
Special Education										
Students with a 504 Plan										
Title I (optional)										
Tested with Standard Accommodation										
Tested with Non-Standard Accommodation										
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report								
Migrant										
Gifted/Talented										
LEP/ELL										
Former LEP Student										
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students								
Free/Reduced Lunch										
Significant Cognitive Disability										
Special Education Disability(ies):										
Autism										
Cognitive Delay										
Deaf-Blindness Impairment										
Deafness										
Emotional Disturbance										
Hearing Impairment										
Learning Disability										
Other Health Impairment										
Orthopedic Impairment										
Speech/Language										
Traumatic Brain Injury										
Visual Impairment										

Data not available for the 2007 report

*Less than ten (10) students were assessed

MontCAS, Phase 2 CRT

Confidential

Mathematics **System Summary Report**

System:
Grade: 03
Spring 2007

III. Results for Subgroups of Students

Reporting Category	System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students										
Gender										
Male										
Female										
Ethnicity										
American Indian or Alaska Native										
Asian										
Hispanic										
Black or African American										
Native Hawaiian or Other Pacific Islander										
White										
Special Education										
Students with a 504 Plan										
Title I (optional)										
Tested with Standard Accommodation										
Tested with Non-Standard Accommodation										
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report								
Migrant										
Gifted/Talented										
LEP/ELL										
Former LEP Student										
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students								
Free/Reduced Lunch										
Significant Cognitive Disability										
Special Education Disability(ies):										
Autism										
Cognitive Delay										
Deaf-Blindness Impairment										
Deafness										
Emotional Disturbance										
Hearing Impairment										
Learning Disability										
Other Health Impairment										
Orthopedic Impairment										
Speech/Language										
Traumatic Brain Injury										
Visual Impairment										

Data not available for the 2007 report

*Less than ten (10) students were assessed

MontCAS, Phase 2 CRT

Confidential

Reading

School Summary Report

School:
System:
Grade: 03
Spring 2007

III. Results for Subgroups of Students

Reporting Category	School					System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students															
Gender															
Male															
Female															
Ethnicity															
American Indian or Alaska Native															
Asian															
Hispanic															
Black or African American															
Native Hawaiian or Other Pacific Islander															
White															
Special Education															
Students with a 504 Plan															
Title I (optional)															
Tested with Standard Accommodation															
Tested with Non-Standard Accommodation															
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report													
Migrant															
Gifted/Talented															
LEP/ELL															
Former LEP Student															
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students													
Free/Reduced Lunch															
Significant Cognitive Disability															
Special Education Disability(ies):															
Autism															
Cognitive Delay															
Deaf-Blindness Impairment															
Deafness															
Emotional Disturbance															
Hearing Impairment															
Learning Disability															
Other Health Impairment															
Orthopedic Impairment															
Speech/Language															
Traumatic Brain Injury															
Visual Impairment															

Data not available for the 2007 report

*Less than ten (10) students were assessed

MontCAS, Phase 2 CRT

Confidential

Mathematics

School Summary Report

School:
System:
Grade: 03
Spring 2007

III. Results for Subgroups of Students

Reporting Category	School					System					State				
	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A	Number	% in N	% in NP	% in P	% in A
All Students															
Gender															
Male															
Female															
Ethnicity															
American Indian or Alaska Native															
Asian															
Hispanic															
Black or African American															
Native Hawaiian or Other Pacific Islander															
White															
Special Education															
Students with a 504 Plan															
Title I (optional)															
Tested with Standard Accommodation															
Tested with Non-Standard Accommodation															
Alternate Assessment		If a student in your system or school took the CRT-Alternate, please refer to Table III on the CRT-Alternate System or School Summary Report													
Migrant															
Gifted/Talented															
LEP/ELL															
Former LEP Student															
LEP Student Enrolled for First Time in a U.S. School		Performance levels are not reported for 1st year LEP students													
Free/Reduced Lunch															
Significant Cognitive Disability															
Special Education Disability(ies):															
Autism															
Cognitive Delay															
Deaf-Blindness Impairment															
Deafness															
Emotional Disturbance															
Hearing Impairment															
Learning Disability															
Other Health Impairment															
Orthopedic Impairment															
Speech/Language															
Traumatic Brain Injury															
Visual Impairment															

Data not available for the 2007 report

*Less than ten (10) students were assessed

APPENDIX E: REPORTING DECISION RULES

Analysis and Reporting Decision Rules
Montana Comprehensive Assessment System (MontCAS) CRT and CRT-Alternate
Spring 06-07 Administration

This document details rules for analysis and reporting. The final student level data set used for analysis and reporting is described in the "Data Processing Specifications." This document is considered a draft until the Montana Office of Public Instruction (OPI) signs off. If there are rules that need to be added or modified after said sign-off, OPI sign off will be obtained for each rule. Details of these additions and modifications will be in the Addendum section.

I. General Information

A. Tests Administered

Grade	Subject	Items included in Raw Score		IABS Reporting Categories (Standards) (Not Applicable for CRT-Alternate)
		CRT	CRT-Alt	
03	Reading Math	Common	All	Cat3
04	Reading Math	Common	All	Cat3
	Science Pilot*	All	N/A	N/A
05	Reading Math	Common	All	Cat3
06	Reading Math	Common	All	Cat3
07	Reading Math	Common	All	Cat3
08	Reading Math	Common	All	Cat3
	Science Pilot*	All	N/A	N/A
10	Reading Math	Common	All	Cat3
	Science Pilot*	All	N/A	N/A

*Pilot administered only to general assessment students.

B. Reports Produced

1. Student Labels
2. Student Report
3. Roster & Item Level Report(online system)

- by grade, subject and class

4. Summary Report
   Consists of sections:
   I. Distribution of Scores
   II. Subtest Results
   III. Results for Subgroups of Students
   - by grade, subject and school
   - by grade, subject and system
   - by grade, subject (state level)

C. Files Produced(excel file format)
   1. One state file for each grade
      a. Consists of student level results
      b. Alternately assessed students are in separate files by grade.

D. School Type

Schtype	**Source**	**Description**	**Included in Aggregations**		
			School	**System**	**State**
"Pras"	Data file provided by state	Private Accredited School. They are their own system	Yes. Same information for school & system but both sets of reports produced	Yes. Same information for school & system but both sets of reports produced	No
"Prnas"	Scanned data	Private non-accredited school. They are their own system	Yes. Same information for school & system but both sets of reports produced	Yes. Same information for school & system but both sets of reports produced	No
"Prnat1"	Scanned data	Private non-accredited Title 1 school. They are their own system.	Yes. Same information for school & system but both sets of reports produced	Yes. Same information for school & system but both sets of reports produced	No
"Oth"	Data file provided by state/Scanned data	non-private school	Yes	Yes	Yes

E. Other Information

1. CRT Tests are constructed with a combination of common and embedded field test items.
2. The CRT-Alternate consists of a set of performance tasks. At grades 3, 5, 6, and 7 the tasks are grouped into five (5) sets of five (5) tasklets for each subject. At grades 4, 8 and 10 the tasks are not grouped.

II. Student Participation/Exclusions

A. Test Attempt Rules

1. A valid response to a multiple choice item is A, B, C, or D. An asterisk (multiple marks) is not considered a valid response.
2. Incomplete (CRT): The student has fewer than two (2) valid responses to common multiple choice items.
3. Incomplete (CRT-Alternate): The student responded to fewer than three (3) items.

B. Not Tested Reasons

N/A

C. Student Participation Status

1. The following students are excluded from all aggregations.
    a. Foreign Exchange Students (FXS).
    b. Home schooled students (SNE).
    c. Part-time students (PSNE).

2. If any of the non-standard accommodations are bubbled the student is considered tested with non-standard accommodations (NSA) in that subject.
3. If the student has First year LEP bubbled and is not Native American the student is considered first year LEP and is excluded from all aggregations.
4. If the student has not been in that school for the entire academic year the student is excluded from school level aggregations (NSAY).
5. If the student has not been in that system for the entire academic year the student is excluded from system level aggregations (NDAY).
6. If the student took the alternate assessment the student is not counted as participating in the general assessment. Alternate Assessment students receive their results on an Alternate Assessment Student Report. They are reported according to participation rules state in this document.
7. (CRT-Alternate) If the teacher halted the administration of the assessment after the student scored zero (0) for three (3) consecutive items (within tasklets for grades 3, 5, 6, and 7) the student is classified as Halted. Scores received after three (3) consecutive zeroes are blanked out and are not counted toward the student's score.

D. Student Participation Summary

Participation Status	Part. Flag	Raw score	Scaled Score	Perf. level	Included on Roster	Included in aggregations		
						Sch	Sys	Sta
FXS	A	Yes	Yes	Yes	No	No	No	No
SNE	A	Yes	Yes	Yes	No	No	No	No
PSNE	A	Yes	Yes	Yes	No	No	No	No

Participation Status	Part. Flag	Raw score	Scaled Score	Perf. level	Included on Roster	Included in aggregations		
						Sch	Sys	Sta
NSA(by subject)	A	Yes	Yes	Yes	Yes	No	No	No
First year LEP	A	Yes	See Report Specific Rules	See Report Specific Rules	Yes	Only in count of First year LEP		
NSAY only	B	Yes	Yes	Yes	Yes	No	Yes	Yes
NDAY	C	Yes	Yes	Yes	Yes	No	No	Yes
ALT*	A	Yes	Yes	Yes	Yes	See footnote below		
Incomplete	A	Yes	Yes	Yes	Yes	No	No	No
Halted(CRT-Alt only by subject)	D	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tested	Z	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* Alternate assessment students are included only in the count of alternate assessment students in general assessment reports. They are included in summary data only for alternate assessment reports (according to participation rules).

III. Calculations

A. Raw Scores
   1. Raw scores are calculated using the scores on common multiple choice and open response items.
   2. Percentages and averages are reported to the nearest whole number.
   3. The number of included students (N) in a subject is the number of students in the school/system/state minus FXS minus PRAS minus PRNAT1 minus PRNAS minus PSNE minus SNE minus First year LEP minus Incomplete minus NSA.
   4. School/system reports are produced regardless of N-size.

B. Scaling

   Scaling is done using constants from psychometrics and the student's raw score.

C. Performance levels are assigned based on the student's earned raw score.

D. Performance Level coding:

Numeric Performance Level	Performance level Name	Abbreviation
1(lowest)	Novice	N
2	Nearing Proficient	NP
3	Proficient	P
4(highest)	Advanced	A

IV. Report Specific Rules

A. Student Label
   1. If a student is First year LEP and incomplete in Reading, the Reading performance level is 'LEP'. The reading scaled score is blank.
   2. If a student is First year LEP, the math performance level is the name of the earned performance level and the scaled score is the student's earned score.
   3. If the student is not first year LEP, the performance level name corresponding to the student's earned score is displayed.
   4. If the student is First year LEP but is not incomplete in Reading then the student receives his earned scaled score and performance level.

B. Student Report
   1. If a student is First year LEP and incomplete in Reading the Reading performance level is 'LEP' and the scaled score is blank.
   2. If the student is First year LEP but is not incomplete in Reading then the student receives his earned scaled score and performance level.
   3. If a student is First year LEP, the math performance level is the name of the earned performance level and the scaled score is the student's earned score.
   4. If the student is not first year LEP, the performance level name corresponding to the student's earned score is displayed.
   5. If the student is incomplete the student receives the scores with a footnote (†) "Student did not complete the assessment."
   6. If the student is NSA the student will receive his scores with the footnote (§) "Student took non-standard accommodation."
   7. There is no last name or first name for the student, the name displayed is "Name Not Provided".
   8. Alt students who are halted receive their scores and performance level and a footnote(§)
      a. Grades 4,8,10 "Teacher halted the administration of the assessment after the student scored a 0 for three consecutive items on different test administrations"
      b. Grades 3,5,6,7 "Teacher halted the administration of one or more of the five test activities after the student scored a 0 for three consecutive items within an activity on two different test administrations. Any completed test activities have been scored and are reflected in the student's scaled score."

C. Roster & Item Level Report
   1. If a student is First year LEP and the student is not incomplete in Reading:
      a. The math performance level is the abbreviation of the earned performance level and the scaled score is the student's earned score.
      b. The reading performance level is the abbreviation of the earned performance level and the scaled score is the student's earned score.
      c. The student is excluded from both Reading and Math aggregations.
   2. If the student is First year LEP and incomplete in Reading
      a. The student's Reading and Math performance levels are 'LEP'.
      b. The student's math and reading earned scaled scores are reported.
      c. The student's responses for both subjects are displayed.
      d. The student is excluded from both math and reading aggregations.
   3. If the student is not first year LEP, the performance level abbreviation corresponding to the student's earned score is displayed.
   4. If the student is incomplete the student receives the scores with a footnote (†) "Student did not complete the assessment."

5. If the student is NSA the student will receive his scores with the footnote (§) "Student took non-standard accommodation."
6. There is no last name or first name for the student, the name displayed is "Name Not Provided".
7. If teacher information is missing the roster is done at the school level.
8. Alternate Assessment students are reported only on their class/school's alternate *Roster & Item Level Report.*

D. School Summary

1. Section III (Results for Subgroups of Students)
    a. Performance level results for subgroups with N less than 10 are suppressed. N is always reported. Footnote * 'Less than 10 students were assessed.'
    b. Count of students who are considered NSA for that subject excluding those students who are incomplete, nsay (at school level), nday (at school and system level) or FXS or SNE or PSNE or First year LEP or alt(general assessment report).

    c. Count of students who are alt excludes those students who are nsay (at school level), nday (at school or system level) or incomplete or FXS or SNE or PSNE or NSA or First year LEP.

    d. Count of First year LEP students excludes those students who are nsay (at school level), nday (at school or system level) or incomplete or FXS or SNE or PSNE or NSA or First year LEP or alt (general assessment).

V. Data File Rules(Excel format**)**

1. The following students are not included in the state file
    a. Alternate Assessment students
    b. Home schooled students(SNE)
    c. Part-Time students (PSNE)
2. If the student receives a performance level 'LEP' on the student report, the student receives LEP for the performance level in the state file.
3. Alt students who are halted are marked '1' in the halted field for that subject.
4. File naming convention:
    a. Studentdatafile[2 digit grade].xls (CRT files)
    b. altStudentdatafile[2 digit grade].xls (CRT-Alternate files)

Accommodations Selection Guidance

Standard Accommodations

Standard accommodations are available to all students on the basis of individual need regardless of disability status. Decisions regarding standard accommodations should be made informally by the student's educational team on an individual basis, consistent with either previous accommodation decisions for the student or current educational needs. Making accommodations decisions on a group basis rather than on an individual basis is not permitted. Any accommodation(s) must be consistent with those used during the student's regular classroom instruction and assessment 2-3 months prior to testing.

Nonstandard Accommodations

If a student uses an accommodation that results in an invalid score (aka, a nonstandard accommodation), the student is considered to be a non-participant when calculating the participation rate for AYP purposes. In addition to counting that student as a non-participant, the score from the assessment is not included in calculating the proficiency rate for AYP determinations.

- Nonstandard accommodations can only be provided for a student with disabilities if the accommodation(s) is specified in the student's IEP.
- If the student is administered the test with a nonstandard accommodation in the content area test (reading, math, or science), the student will not be counted as a participant for AYP determinations in that content area. The nonstandard accommodation used must be coded in the appropriate box(es) on page two of the Student Response Booklet (SRB). The student's results for that content area test (reading, math, or science) will not be calculated in the averages for AYP determination.
- The Individuals with Disabilities Education Act (IDEA) requires that all students participate in the statewide assessment. This requirement applies whether or not the student takes the test with a nonstandard accommodation.

Type of Accommodation	**ELL Students**	
Scheduling Accommodations	Direct	Indirect
1. Change in Administration Time: Test is administered at a time of day or a day of the week based on student needs.		
2. Session Duration: Test is administered in appropriate blocks of time for individual student needs, followed by rest breaks.		
3. Extended Time: Time is extended beyond the regular test administration allotments until, in the administrator's judgment, the student could no longer sustain the activity.	X	x

Type of Accommodation	ELL Students	
Setting Accommodations	Direct	Indirect
4. Individual Administration: Test was administered in a one to one situation.		x
5. Small Group Administration: Test was administered to a small group of students.		x
6. Reduce Distractors: Student is seated at a carrel or other physical arrangement that reduces visual distraction.		
7. Alternative Setting: Test is administered to the student in a different setting.		x
8. Change in Personnel: Test is administered by other personnel known to the student (e.g., LEP, Title I, special education teacher).	X	
9. Home Setting: Test is administered to the student by school personnel in their home.		
10. Front Row Seating: A student is seated in front of the classroom when taking the test.	X	
11. Teacher Presence: A teacher faces the student during test administration.		
Equipment Accommodations	Direct	Indirect
12. Magnification: Student used equipment to magnify test materials.		
13. Noise Buffers: Student wears equipment to reduce environmental noises.		
14. Template: Student uses a template.		
15. Amplification: Student uses amplification equipment (e.g., hearing aid or auditory trainer) while taking test.		
16. Writing Tools: Student uses a typewriter or word processor (without activating spellchecker).		
17. Voice Activation: Student speaks response into computer equipped with voice activation software.		
18. Bilingual Dictionary: Student uses a bilingual dictionary (Note: Bilingual dictionary could include a simplified English dictionary or glossary, subject area vocabulary list).	X	
Recording Accommodations	Direct	Indirect
19. Dictation: The student dictates answers to a test administrator who records them in the Test Booklet.		
20. Writing Tools: The student marks or writes answers with the assistance of a technology device or special equipment. The students' answers are transferred by the test administrator to the Test Booklet.		
21. Assistive Technology: Another form of assistive technology routinely used by the student (that does not change the intent or content of the test) was used by the student.		

Type of Accommodation	ELL Students	
Modality Accommodations	Direct	Indirect
22. Oral Presentation: Tests were read to the student by the test administrator (with the exception of reading passages). Note: Readers must read test items/questions to the student word-for-word exactly as written. Readers may not clarify, elaborate, or provide assistance to the student regarding the meaning of words, intent of test questions, or responses to test items/questions.	X	
23. Test Interpretation: Tests, including directions, were interpreted for students who are deaf or hearing-impaired (with the exception of interpreting the reading test).		
24. Test Directions with Verification: An administrator gave test directions with verification (by using a highlighter) that the student understood them.	X	
25. Test Directions Support: An administrator assisted students in understanding test directions, including giving directions in native language.	X	
26. Sheltered English: Test was read to an LEP student in "sheltered English" (with the exception of reading the reading test).	X	
27. Braille: A braille version of the test was used by the student.		
28. Large Print: A large print version of the test was used by the student.		
29. Other: With verification from OPI in advance of the testing window, some other approved accommodation was used by a student.	X	

APPENDIX F: SUBGROUP RELIABILITIES

Table F-1. Reliabilities of Subgroups by Grade and Subject.

Grade	Subject	Subgroup	N	(α)
3	Math	White	8555	0.89
		Native Hawaiian or Pacific Islander	19	0.85
		Hispanic or Latino	261	0.88
		Black or African American	117	0.89
		Asian	93	0.87
		American Indian or Alaskan Native	1227	0.90
		LEP	487	0.89
		IEP	1181	0.91
		Low SES	4173	0.89
	Reading	White	8527	0.90
		Native Hawaiian or Pacific Islander	19	0.86
		Hispanic or Latino	257	0.89
		Black or African American	116	0.89
		Asian	92	0.90
		American Indian or Alaskan Native	1217	0.89
		LEP	480	0.88
		IEP	1138	0.91
		Low SES	4146	0.90
4	Math	White	8454	0.90
		Native Hawaiian or Pacific Islander	30	0.91
		Hispanic or Latino	294	0.90
		Black or African American	118	0.91
		Asian	79	0.92
		American Indian or Alaskan Native	1193	0.91
		LEP	432	0.89
		IEP	1157	0.91
		Low SES	4060	0.91
	Reading	White	8428	0.89
		Native Hawaiian or Pacific Islander	29	0.91
		Hispanic or Latino	293	0.89
		Black or African American	118	0.87
		Asian	77	0.90
		American Indian or Alaskan Native	1187	0.88
		LEP	426	0.85
		IEP	1124	0.90
		Low SES	4036	0.89

Table F-1. Reliabilities of Subgroups by Grade and Subject

Grade	Subject	Subgroup	N	(α)
5	Math	White	8776	0.90
		Native Hawaiian or Pacific Islander	25	0.92
		Hispanic or Latino	286	0.91
		Black or African American	102	0.91
		Asian	114	0.92
		American Indian or Alaskan Native	1183	0.90
		LEP	387	0.88
		IEP	1215	0.89
		Low SES	4029	0.90
	Reading	White	8759	0.89
		Native Hawaiian or Pacific Islander	25	0.82
		Hispanic or Latino	283	0.90
		Black or African American	101	0.87
		Asian	114	0.90
		American Indian or Alaskan Native	1182	0.89
		LEP	385	0.87
		IEP	1194	0.89
		Low SES	4019	0.89

Table F-1. Reliabilities of Subgroups by Grade and Subject

Grade	Subject	Subgroup	N	(α)
6	Math	White	8890	0.90
		Native Hawaiian or Pacific Islander	29	0.91
		Hispanic or Latino	247	0.90
		Black or African American	91	0.86
		Asian	104	0.91
		American Indian or Alaskan Native	1158	0.88
		LEP	421	0.84
		IEP	1145	0.87
		Low SES	3908	0.88
	Reading	White	8879	0.88
		Native Hawaiian or Pacific Islander	29	0.81
		Hispanic or Latino	244	0.87
		Black or African American	91	0.86
		Asian	104	0.88
		American Indian or Alaskan Native	1156	0.88
		LEP	416	0.82
		IEP	1128	0.88
		Low SES	3899	0.89
7	Math	White	9231	0.89
		Native Hawaiian or Pacific Islander	32	0.84
		Hispanic or Latino	257	0.89
		Black or African American	121	0.88
		Asian	102	0.89
		American Indian or Alaskan Native	1194	0.86
		LEP	472	0.78
		IEP	1273	0.83
		Low SES	4037	0.88
	Reading	White	9225	0.89
		Native Hawaiian or Pacific Islander	32	0.86
		Hispanic or Latino	256	0.89
		Black or African American	121	0.88
		Asian	101	0.85
		American Indian or Alaskan Native	1198	0.90
		LEP	467	0.85
		IEP	1270	0.87
		Low SES	4027	0.90

Table F-1. Reliabilities of Subgroups by Grade and Subject

Grade	Subject	Subgroup	N	(α)
8	Math	White	9432	0.92
		Native Hawaiian or Pacific Islander	18	0.90
		Hispanic or Latino	246	0.90
		Black or African American	81	0.91
		Asian	91	0.92
		American Indian or Alaskan Native	1208	0.90
		LEP	487	0.82
		IEP	1293	0.86
		Low SES	3835	0.91
	Reading	White	9436	0.90
		Native Hawaiian or Pacific Islander	18	0.89
		Hispanic or Latino	244	0.89
		Black or African American	80	0.85
		Asian	90	0.88
		American Indian or Alaskan Native	1213	0.91
		LEP	482	0.87
		IEP	1299	0.89
		Low SES	3834	0.91
10	Math	White	9682	0.91
		Native Hawaiian or Pacific Islander	28	0.88
		Hispanic or Latino	218	0.89
		Black or African American	76	0.90
		Asian	107	0.92
		American Indian or Alaskan Native	1045	0.88
		LEP	377	0.80
		IEP	1060	0.83
		Low SES	2991	0.90
	Reading	White	9688	0.89
		Native Hawaiian or Pacific Islander	28	0.90
		Hispanic or Latino	218	0.90
		Black or African American	74	0.90
		Asian	107	0.89
		American Indian or Alaskan Native	1051	0.90
		LEP	381	0.86
		IEP	1058	0.89
		Low SES	2996	0.91

[1]Only subgroups with sample size ≥10 reported